Quantitative Monitoring of Aerobic and Anaerobic Bioprocesses using Vibrational Spectroscopy

John Dahlbacka

PhD Thesis in Process Design and Systems Engineering
Faculty of Science and Engineering
Åbo Akademi University

Åbo, Finland 2018

Quantitative Monitoring of Aerobic and Anaerobic Bioprocesses using Vibrational Spectroscopy

John Dahlbacka

PhD Thesis in Process Design and Systems Engineering
Faculty of Science and Engineering
Åbo Akademi University

Åbo, Finland 2018

Dissertations published by Process Design and Systems Engineering
ISSN 2489-7272
978-952-12-3682-2
978-952-12-3683-9 (pdf)
Painosalama Oy
Åbo 2018

Preface

The journey towards the completion of this thesis started several years ago at the Process Design and Systems Engineering Laboratory at the Faculty of Science and Engineering at Åbo Akademi University (at that time known as the Process Design Laboratory at the Faculty of Chemical Engineering). Many things have changed since that time, but one fact remains: Being able to perform quantitative measurements in bioprocesses using vibrational spectroscopy is still of great interest. However, doing so requires creation of reliable calibration models. Construction of reliable calibration models to be used in the context of multi-constituent slurries is, in turn, a challenge. Therefore, this thesis focuses not only on the usefulness of the measurement applications themselves, but also on (perhaps somewhat unorthodox) methods, chemical as well as mathematical, that can either shorten the time spent on creating a calibration data set and/or produce more accurate calibration models.

I want to express my gratitude to my supervisor Professor Anders Brink and my assisting supervisor Professor Henrik Saxén. Special thanks also go to my former supervisor and now retired Professor Kaj Fagervik, especially for keeping a positive attitude throughout this process. Former and present colleagues have also played important roles by providing a stimulating environment for me to work in; none mentioned, none forgotten. Finally, yet importantly, I thank my wife Lakkana for all the things in life that are far more important than doctoral dissertations.

Vasa, October 2017, John Dahlbacka

Abstract

In bioprocesses, living cells are used to produce biological materials of interest. Perhaps the most well-known product produced in bioprocesses is penicillin. Implementing process analytical technologies (PAT) in bioprocesses in terms of quantitative measurements using vibrational spectroscopy is important, but also challenging. In comparison to chemical processes in general, bioprocesses can be described as very complex, where numerous constituents play a vital role within the cells as well as outside the cells. Although bioprocess engineering relies on utilising the natural or genetically modified behaviour of the cultivated organism or organisms, what is optimal for the well-being of the organism is not necessarily optimal for the production of the compound of interest. Therefore, the organisms' environment needs to be monitored and controlled. In bioreactors, this is a fairly straightforward task when it comes to parameters such as pressure, temperature, dissolved oxygen and pH. However, this is not the case when it comes to the important chemical composition of the organisms' environment, or for that matter, the intracellular constituents. For monitoring the chemical environment, vibrational spectroscopy is a very versatile and thereby very interesting method.

Quantitative measurements in bioprocesses using mid or near infrared spectroscopy typically require multivariate calibration methods and a significant amount of calibration data on which to base the model. Collecting calibration data is, in turn, usually a very tedious process. However, even when a sufficient amount of calibration data is available and a multivariate method is used, there is no guarantee that the obtained accuracy is sufficient for the intended measurement application. Therefore, two challenges for implementing quantitative measurements in bioprocesses can be easily identified: (1) How can the time spent on collecting calibration data be reduced, and (2) What mathematical methods can be used to increase the accuracy of the calibration model? These questions are also addressed in this thesis. This is done by evaluating the impact of the methodology used in applications of quantitative measurements on *Pichia pastoris*, *Streptomyces peucetius*, and anaerobic digestion processes. In general, the methodologies used produced promising results.

Sammanfattning

I bioprocesser används levande celler för att producera biologiskt material av mycket varierande slag. Den kanske mest välkända produkten från bioprocesser är penicillin. För att effektivt kunna utnyttja bioprocesser är det väldigt viktigt med en fungerande processanalytik. En form av processanalytik som bedöms vara av stort intresse är kvantitativa mätningar baserade på vibrations-spektroskopi. Denna omfattar tre olika tekniker, nära-infraröd-, medium-infraröd- och Ramanspektroskopi. Gemensamt för dessa tekniker är att de ger information om vibrations- och rotationsövergångar hos molekyler. De infraröda teknikerna förutsätter att vibrationerna medför en förändring av molekylens dipolmoment, medan vibrationer som leder till en förändring av den molekylära polariserbarheten är Raman-aktiva. Denna avhandling omfattar endast de infraröda teknikerna. Förenklat kan funktionen av dessa beskrivas med att då ljus träffar en molekyl så kan det uppstå vibrationer i molekylbindningarna. Energin i vibrationerna är relaterad till energin hos det ljus som orsakar dem. Genom att studera vilka våglängder som försvagats då ljuset varit i kontakt med provet fås därigenom information om vilka molekylbindningar som finns i provet. Hur mycket en våglängd försvagats ger i sin tur information om hur många molekylbindningar det infallande ljuset kommit i kontakt med. I praktiken innebär detta att de infraröda teknikerna kan användas för både kvalitativ och kvantitativ analys.

Bioprocesser baserar sig på kemiska reaktioner orsakade av mikrobiell aktivitet. I jämförelse med typiska kemiska processer är de väldigt komplexa och förutsätter samt genererar ett stort antal viktiga kemiska komponenter. Detta gör det svårt att kartlägga, monitorera och reglera bioprocesser. Utöver de utmaningar som komplexiteten i sig själv för med sig är den metaboliska aktivitet som är optimal för själva organismen inte nödvändigtvis optimal för produktionen av den önskade substansen. En förutsättning för att kunna bedriva produktion baserad på bioprocesser är därmed att organismernas omgivning kan monitoreras och därigenom kontrolleras. Vissa parametrar, såsom tryck, temperatur, halten löst syre och pH, är förhållandevis lätta att både mäta och reglera. Detta är i regel inte fallet då det gäller den kemiska sammansättningen av organismernas

omgivning, för att inte tala om förekomsten av metaboliska mellan- och slutprodukter inuti cellerna. Vad gäller åtminstone bestämning av den kemiska sammansättningen av organismernas omgivning så utgör vibrationsspektroskopi ett väldigt flexibelt och därigenom intressant alternativ.

Kvantitativ monitorering av bioprocesser med hjälp av nära- och medium-infrarödspektroskopi förutsätter i regel att en betydande mängd kalibreringsdata finns tillgänglig och att kalibreringsmodellen baserar sig på multivariata metoder. Behovet av en betydande mängd kalibreringsdata medför i sin tur att stora ekonomiska och tidsmässiga resurser måste avsättas för att åstadkomma en kalibrering. Tyvärr medför användandet av multivariata metoder inte heller alltid att mätnoggrannheten blir tillräcklig för en planerad mätapplikation, trots att en tillräcklig mängd kalibreringsdata finns till förfogande. Ur dessa två påståenden kan extraheras två mycket relevanta frågeställning kring användandet av vibrationsspektroskopi för kvantitativa mätningar i bioprocesser: (1) Hur kan kalibreringsdata erhållas med minimal resursförbrukning och (2) Vilka matematiska metoder kan tillämpas för att göra kalibreringsmodellen noggrannare?

Kvantitativa modeller för monitorering av bioprocesser med vibrationsspektroskopi baserar sig vanligen på spektra tagna på (a) vattenlösningar på enskilda rena komponenter, (b) vattenlösningar med en blandning av rena komponenter, eller (c) obehandlade eller behandlade processprov. Även om (a) medför att kalibreringsdata snabbt kan skapas är tillvägagångssättet tveksamt eftersom verkliga processprov innehåller en mängd andra kända och okända komponenter för vilka kalibreringsmodellen inte kan vara robust. Alternativ (b) är något bättre och kan eventuellt fungera förutsatt att merparten av den spektrala informationen i processproven kan förklaras med de ingående komponenterna. Spontant förefaller då kanske alternativ (c) som det bästa och i praktiken är det väl också ofta så, men detta alternativ för också med sig ett antal nackdelar. För det första så begränsas den vibrationsspektroskopiska mätningens noggrannhet av referensmetodens noggrannhet, för det andra kan det vara väldigt resursförbrukande att producera prov och utföra referensmätningar och för det tredje går det inte att hantera korskorrelationen mellan olika komponenter i denna typ av prov. Det sistnämnda är ett stor problem för satsvisa bioprocesser i synnerhet, eftersom i stort sett allting korrelerar med allting i dessa. Som alternativa sätt att skapa kalibreringsdata har därför i denna avhandling undersökts potentialen i att använda syntetiska multikomponentspektra baserade på spektra av rena komponenter, semisyntetiska processprovsspektra baserade på processprovsspektra och spektra av rena komponenter, samt spetsning av processprov med rena komponenter.

I stort sett i alla kvantitativa applikationer av medium-infrarödspektroskopi och det stora flertalet applikationer av nära-infrarödspektroskopi i bioprocessammanhang tillämpas den partiella minstakvadrat-metoden för kalibreringsändamål. Det har framförts att den är onödigt avancerad då medium-infrarödspektroskopi utgör mätmetoden, men samtidigt finns många exempel på förbättrad mätnoggrannhet för kvantitativ nära-infrarödspektroskopi då prediktionen baserar sig på lokala modeller. Dock inte för bioprocess-tillämpningar i någon betydande omfattning. I denna avhandling utvärderas därför också tillämpningen av en lokal iterativ kalibreringsmetod för kvantitativa applikationer med nära-infrarödspektroskopi.

I denna avhandling testades användning av syntetiska multikomponentspektra som kalibreringsdata i fermenteringar av *Pichia pastoris* och *Streptomyces peucetius* med relativt god framgång. I *Pichia pastoris* tillämpningen användes också semisyntetiska processprovsspektra med ett mycket gott resultat. Spetsade processprov testades i en mätapplikation för rötningsprocessen med viss, men samtidigt kanske inte obestridbar, framgång. Hur väl lokala kalibreringsmetoder lämpar sig i det sammanhang som denna avhandling utgör testades också på rötningsprocessen. Resultatet var riktigt bra för en komponent och tvetydigt för en annan och utfallet därmed icke entydigt.

Publications

The author of this thesis has been involved in research related to vibrational spectroscopy and bioprocess engineering for a number of years. This has resulted in, among other things, eleven peer review publications and five publications in conference proceedings. Among the peer-review publications, five contain the combination of vibrational spectroscopy and bioprocesses; these are therefore included in this thesis. Thus, this thesis is to a large extent based on the five publications specified below, although some additional results and information are also included. These publications will be referred to using Roman numerals from this point forward.

Included in this thesis

I. Dahlbacka, J., Weegar, J., von Weymarn, N., & Fagervik, K. (2012). On-line measurement of the substrate concentrations in *Pichia pastoris* fermentations using FT-IR/ATR. *Biotechnology Letters*, *34*(6), 1009-1017.

II. Dahlbacka, J., Kiviharju, K., Eerikäinen, T., & Fagervik, K. (2013). Monitoring of *Streptomyces peucetius* cultivations using FTIR/ATR spectroscopy and quantitative models based on library type data. *Biotechnology Letters*, *35*(3), 337-343.

III. Dahlbacka, J., Lillhonga, T., & Döring, M. (2013). Designed orthogonal sample spiking based calibrations for quantitative liquid phase measurements with near infrared spectroscopy in an anaerobic digestion process. *Journal of Near Infrared Spectroscopy*, *21*(1), 11-22.

IV. Dahlbacka, J., & Lillhonga, T. (2013). Quantitative measurements of anaerobic digestion process parameters using near infrared spectroscopy and local calibration models. *Journal of Near Infrared Spectroscopy*, *21*(1), 23-33.

V. Dahlbacka, J., Pohar, J., & Lillhonga, T. (2014.) Some near infrared spectroscopy applications of an iterative calibration model

regression strategy – A proof of concept study. *Journal of Near Infrared Spectroscopy*, *22*(6), 389-400.

Contributions by the author

The author of this thesis, denoted "the Author" below, is the main contributor and first author to all of the included publications. A somewhat more detailed description of the contribution is given below.

I. The Author participated in carrying out the fermentations, in the on-line collection of spectra and in the reference measurements. Everything related to multivariate calibration model building, evaluation, and calibration methodology, as well as off-line collection of spectra, can be attributed to the Author. The manuscript was written in whole by the Author.

II. The Author carried out the fermentations from which the spectra were collected, collected the spectra and performed most of the reference measurements. Everything related to multivariate calibration model building, evaluation, and calibration methodology, as well as off-line collection of spectra, can be attributed to the Author. The manuscript was written in whole by the Author.

III. The Author made a smaller contribution in the fermentations and the implementation of the spiking protocol. All measurements on centrifuged samples were done by the Author. The concept of using an orthogonal spiking procedure in this application, the spiking protocol as such and all calibration models presented can be attributed to the Author. The manuscript was written in whole by the Author.

IV. The methodology presented can mainly be attributed to the Author. The script used was written by the Author and all modelling was carried out by the Author. The manuscript was written in whole by the Author.

V. The methodology used in ML-PLS and the calibration results presented is attributed to the Author. The manuscript was written in whole by the Author.

List of abbreviations

AD	anaerobic digestion
adiff	absolute distance in spectral space
adpca	absolute distance in PC score space
adpls	absolute distance in PLS score space
ANN	artificial neural network
AOX1	alcohol oxidase I gene
ATR	attenuated total reflectance
BANN	Bayesian artificial neural network
BRT	boosted regression trees
cdiff	correlation in spectral space
cdpls	correlation in PLS score space
DiComp	diamond composite
ediff	Euclidean distance in spectral space
edpca	Euclidean distance in PC score space
edpls	Euclidean distance in PLS score space
epred	Euclidean distance to prediction
FS	fermentation spectrum/spectrum collected from the fermentation broth
FT-IR	Fourier transform infrared
GPR	Gaussian process regression
GUI	graphical user interface
HPLC	high-performance liquid chromatograph(y)
IR	Infrared
KPLS	kernel partial least squares
LS-SVM	least-squares support vector machine
LWR	locally weighted regression
MCT	mercury cadmium telluride
mdpca	Mahalanobis distance in PC score space
mdpls	Mahalanobis distance in PLS score space
MIR	mid infrared
ML-PLS	multi-layer partial least squares
MLR	multiple linear regression

MMS	monolithic miniature spectrometer
Nef	negative factor
NIR	near infrared
OSC	orthogonal signal correction
PAT	process analytical technologies
PC	principal component
PCR	principal component regression
PGS	plane grating spectrometer
PLS	partial least squares
PM	Pharmamedia (medium)
RER	range error ratio
RMSEC	root mean square error of calibration
RMSECV	root mean square error of cross validation
RMSEP	root mean square error of prediction
RPD	ratio of standard error of performance to standard deviation
rpm	revolutions per minute
RVM	relevance vector machines
SBL	spectrum-based learner
SEP	standard error of prediction
SP	soy peptone (medium)
SS	simulated background matrix spectrum
SVM	support vector machine
SVMR	support vector machine regression
TVFA	total volatile fatty acids
VFA	volatile fatty acids
VVM	volume of air to volume of medium per minute
yres	absolute Y-residual

Contents

Chapter 1

1. Introduction

This thesis studies the usability of near infrared (NIR) and mid infrared (MIR) spectroscopy in the context of quantitative measurements in bioprocesses. These spectroscopic techniques are described in (for instance) Osborne et al. (1993) and Stuart (2005). To avoid confusion, it should be stated that in this thesis, the term infrared (IR) spectroscopy refers to the combination of MIR and NIR spectroscopy, although IR spectroscopy also includes far infrared spectroscopy and the term IR spectroscopy is also commonly used to describe MIR spectroscopy exclusively. It is the author's impression that quantitative measurements in bioprocesses using IR spectroscopy are challenging due to, among other things, the relatively low concentrations of the constituents of interest, the high absorption of IR light caused by water, and the complex sample matrix. These challenges can perhaps be met by the availability of high-quality calibration data and advanced calibration methods. Thus, this thesis also investigates how calibration data can be generated efficiently, and whether local calibration methods can be beneficial in comparison to the traditional partial least squares (PLS) model regression. A detailed description of PLS regression can be found in Geladi & Kowalski (1986).

In this thesis, the usability of IR spectroscopy for quantitative bioprocess monitoring was evaluated in terms of creating measurement applications for fermentations of *Pichia pastoris* and *Streptomyces peucetius*, as well as for the anaerobic digestion (AD) process. In the case of the fermentations of *Pichia pastoris*, measurements of the glycerol and methanol concentration using MIR spectroscopy were carried out. Both the strain X33 (phenotype Mut^{+}) and the strain KM71H (phenotype Mut^{S}) were used in the cultivations. Since on-line information about the substrate is of great interest (as described later on), the research evaluated whether simple mathematical manipulations of the available calibration spectra could be used to rapidly create a calibration data set from which a reliable calibration model could be regressed. In the case of *Streptomyces peucetius* fermentations, measurements of the glucose, starch, and acetate concentrations were implemented by means of MIR spectroscopy. These measurements also relied on calibrations based on mathematical manipulations of the calibration data. In the case of anaerobic digestion, measurement applications were created for ammonium, acetate, propionate and total volatile fatty acids (TVFA). In this application, three questions were addressed: (1) Is spiking of process samples using pure component solutions a feasible method for "fast" production of calibration data? (2) Is it favourable to use MIR in comparison to NIR spectroscopy in this application? (3) Can the use of local calibration methods result in more accurate/reliable calibration models? The spiking of the process samples was carried out as a three-factor central composite design, with axial points at the cube walls and no centre points, yielding 14 calibration samples for each original digester sample. The comparison between NIR and MIR was carried out, to a large extent, by an unsupervised construction and validation of calibration models. The intention was to rule out the human factor by doing thousands of calibrations while following the exact same steps. The local calibration techniques applied here were the locally weighted regression (LWR) and multi-layer PLS (ML-PLS), both of which are described below. The use of local calibration techniques in bioprocess monitoring has probably not been attempted before, and is therefore of special interest for this thesis.

1.1 Near and mid infrared spectroscopy

Both NIR and MIR measurements are based on analysing the absorption of radiation as a function of wavelength by functional group vibrations. These techniques are thereby very similar; at the same time, there are differences between the methods, both in terms of historical background and with regard to the information obtained.

MIR spectroscopy as such emerged at the beginning of the 20th century, and spectrometers have been commercially available since the 1940s. Organic chemistry research and industrial laboratory analysis powered the development of the technology in the 60s and 70s, resulting in the important introduction of the FT-IR (Fourier transform infrared) spectrometer in the 70s. In the 80s, the cost associated with the FT-IR technology became more moderate, making it the standard MIR technique. (Doyle, 1995)

NIR spectroscopy as an analytical technique emerged in the 1960s through the work of Norris and his co-workers (e.g. Hart et al., 1962). Whereas MIR spectroscopy can be said to have a background in laboratory analysis, NIR spectroscopy evolved later and for other reasons. NIR emerged as a technique for solving practical quality control problems, rather than for performing high-resolution molecular structure analysis. In addition, measurements on highly scattering agricultural products, such as grain, require a high throughput technique enabling a large collecting area. These requirements are met by NIR spectroscopy, which is capable of performing diffuse reflectance or diffuse transmittance measurements. Another advantage with NIR is that it can be readily implemented in combination with fibre optics, enabling multiple sampling devices and transmission of signals to and from remote locations. Thus, from an historical perspective, MIR spectroscopy is associated with laboratory measurements, and NIR spectroscopy is associated with process monitoring and quality control. (Doyle, 1995)

MIR and NIR spectroscopy techniques rely on the vibrations of the atoms in a molecule. Depending on the composition of a sample, infrared light passing through it may be absorbed at specific wavelengths. By observing several wavelengths at a time, a spectrum is obtained. The wavelength that is absorbed has an energy corresponding to the frequency of a vibration of a part of a sample molecule. This process will produce peaks in the studied spectrum. In order for an absorption to occur, the electric dipole moment of the molecule must change during the vibration (or rotation). Molecules that fulfil this characteristic are said to be infrared-active. This seems to limit somewhat the use of IR spectroscopy, but it is still reasonable to claim that virtually any sample in virtually any state may be studied. What is referred to here as vibrations is, in fact, the bending and stretching of molecular bonds. For a single molecule, several different modes of vibrations can occur. For instance, in the methylene group, vibrations can occur as rocking, wagging, twisting, scissoring, as well as symmetric and asymmetric stretching. (Stuart, 2004)

MIR covers the wavenumber region between 4000 and 400 cm^{-1} (2500-25000 nm in terms of wavelengths), whereas NIR covers the region between 14000 and 4000 cm^{-1} (~700-2500 nm). The information obtained in the MIR region originates from fundamental vibrations. A fundamental vibration occurs when a quantum of energy corresponding to the vibration's frequency is absorbed. The information in the NIR region derives from overtones and combinations of the fundamental vibrations in the MIR region. The first overtone of a fundamental vibration occurs when two quanta are simultaneously absorbed. The first overtone will therefore appear in the spectrum at twice the wavenumber of the fundamental vibration. Similarly, the second overtone is obtained when three quanta are absorbed. Combination bands, in turn, occur when two or more fundamental bands absorb energy simultaneously. An overtone transition is much less likely to occur than a fundamental one. Therefore, their intensity is typically very weak and decreases by a factor of 10 from one overtone to the next. Furthermore, another difference between MIR and NIR is that the bands in NIR are typically more overlapped than in MIR. (Stuart, 2004)

Since many of the molecules or constituents of interest in a bioprocess are in fact infrared-active, the use of MIR and NIR spectroscopy can be said to provide an abundance of chemical information. This, in combination with multivariate data analysis, enables quantification of individual constituents even in a complex matrix (Cooper et al., 1997). This, in turn, is the typical situation in bioprocess monitoring, where the matrix is complex and multivariate methods are therefore routinely applied. Other arguments for using vibrational spectroscopy typically include the words "fast", "versatile" and "non-invasive". This is all true; however, as will be discussed, implementing quantitative measurement applications using IR spectroscopy (in bioprocesses) also comes with some challenges.

1.2 NIR and MIR spectroscopy in bioprocess monitoring

Several review papers that describe the different applications of NIR and MIR spectroscopy in bioprocess monitoring can be found (e.g. Abu-Absi et al., 2014; Beutel & Henkel, 2011; Biechele et al., 2015; Landgrebe et al., 2010; Lourenço et al., 2012); therefore, no attempt to make a comprehensive listing of these applications is made in this thesis. Instead, the focus is on why and how infrared spectroscopy is used for bioprocess monitoring. Bioprocesses will be of great future economic importance, and the cultivation of microorganisms in bioreactors is a critical unit operation of biotechnological processes (Lourenço et al., 2012). Effective substrate consumption or maximal production, in turn, typically relies on continuous sampling from which the analytical data needed for process control are obtained (Beutel & Henkel, 2011). Therefore, on a very general level, it can be said that improved methods for monitoring bioprocesses will play an important role in the future development of bioprocess technology (Clementschitsch & Bayer, 2006). More specifically, it is the real-time measurement of important process variables that is the basis for process control, exact process documentation, and high productivity (Landgrebe et al., 2010). This requires the ability to measure the variables in-line, which is very common practice for physical parameters such as temperature and pH, but more difficult for metabolites, nutrients and cellular components (Abu-Absi et al., 2014).

Process measurements can be carried out in many ways using many types of sensors. Even in the somewhat narrow context of bioprocess monitoring using IR spectroscopy, a clarification of the terminology used can therefore be useful. According to Biechele et al. (2015):

> "A sensor that is directly interfaced and in contact with the process fluids is called an in-line or in-situ sensor. If the sample is withdrawn via a filtration module and analysed outside the bioreactor, the analysis is termed at-line. If in-line and at-line sensors measure a quantity (e.g., pH value) continuously, the analysis can be regarded on-line. If sensor data are available without any time delay, real-time monitoring is possible. All

other measurements are considered to be off-line. Invasive and non-invasive sensors are classified by the interaction of the sensor and the bioprocess. If the bioprocess medium interacts with the sensor as it is done for biosensors, the sensor is called invasive. If no interaction between sensor and medium occurs, the sensor is non-invasive".

Currently, the chemical components of the bioreactor media are mainly measured off-line (Lourenço et al., 2012). This requires sample withdrawal, which implies a contamination hazard, typically followed by different types of time-consuming sample preparations. In-situ sensors mounted as an integrated part of the bioreactor remove the contamination risk, but this measurement setup also implies that the sensors must endure the conditions during sterilisation (Beutel & Henkel, 2011). Bioreactors can be described as complex multi-variable systems, where substrates are consumed while products and intermediates are formed. The cells are most often suspended in the bioreactor medium and a gas phase is present in the form of bubbles (Bluma et al., 2010). Furthermore, the medium itself may consist of undefined additives, and even a chemically-defined medium can contain more than 50 constituents. Therefore, the sensors used for bioprocess monitoring must be able to function in the presence of a complex multiphasic matrix, and still be able to precisely measure low concentrations of various constituents (Lourenço et al., 2012).

For a sensor, some of the most important parameters can be described as below (as a revised, but otherwise directly quoted, list from Bochenkov & Sergeev, (2010)):

- Sensitivity is a change of measured signal per analyte concentration unit, i.e., the slope of a calibration graph. This parameter is sometimes confused with the detection limit.
- Selectivity refers to characteristics that determine whether a sensor can respond selectively to a group of analytes or even specifically to a single analyte.
- Stability is the ability of a sensor to provide reproducible results for a certain period of time. This includes retaining the sensitivity, selectivity, response, and recovery time.
- Detection limit is the lowest concentration of the analyte that can be detected by the sensor under given conditions, particularly at a given temperature.

- Dynamic range is the analyte concentration range between the detection limit and the highest limiting concentration.
- Linearity is the relative deviation of an experimentally determined calibration graph from an ideal straight line.
- Resolution is the lowest concentration difference that can be distinguished by the sensor.
- Response time is the time required for the sensor to respond to a step concentration change from zero to a certain concentration value.
- Working temperature is usually the temperature that corresponds to maximum sensitivity.

Both Beutel & Henkel (2011) and Biechele et al. (2015) mention the response time as an important parameter. In general, the response time should be small relative to the important process dynamics (Biechele et al., 2015), and fast-growing prokaryotes require a much higher analytical rate than do eukaryotes (Beutel & Henkel, 2011). However, it seems reasonable to suggest that in the context of bioprocess measurement using NIR or MIR spectroscopy, the response time (typically less than a minute) should rarely be an issue. In general, sensor systems for bioprocess monitoring should then provide high selectivity, sensitivity, robustness, repeatability, stability, low detection limit, a short response time and a long life (Biechele et al., 2015), while minimally affected by fouling (Beutel & Henkel, 2011). Furthermore, the ability to measure multiple constituents with the same sensor instrument is certainly beneficial (Abu-Absi et al., 2014).

In-situ measurement systems can be realised by a variety of sensing principles, although spectroscopic techniques are regarded as the most promising due to the broad range of information obtained with them (Beutel & Henkel, 2011). In other words, many of the criteria specified above can be met by spectroscopic methods, and their usefulness in bioprocess applications is increasing (Abu-Absi et al., 2014). In addition, it can be said that most other options are still high-priced, require frequent maintenance, and are normally limited to single-property analysis (Lourenço et al., 2012). When it comes to NIR and MIR spectroscopy, much of the current effort by the instrument industry is put into creating smaller and cheaper instruments; however, they can also still be generally regarded as fairly expensive instruments. Despite this, MIR and NIR sensors have nevertheless received growing attention in recent years (Landgrebe et al., 2010).

This can be attributed to the fact that readily found, commercially-available systems can be easily adapted for bioprocess monitoring, at the same time as these systems allow for simultaneous detection of multiple constituents (Landgrebe et al., 2010). Used as an in-situ measurement, these techniques also require no sampling and no analyte consumption or other reagents, while at the same time being a non-invasive measurement (Beutel & Henkel, 2011).

Most of the scientific publications dealing with IR spectroscopic measurements of bioprocesses deal with NIR rather than MIR measurements, due to the simplicity of the instrumentation involved (Abu-Absi et al., 2014). However, MIR spectroscopy is evidently gaining popularity through improved optics and smaller, more powerful lasers and detectors (Abu-Absi et al., 2014), as well as through the use of optical fibre probes in combination with attenuated total reflectance (ATR) (Lourenço et al., 2012). For in-situ process monitoring in particular, the possibility of using probes coupled to optical fibres is a key feature. In this sense, MIR is still the less versatile technique in comparison to NIR, because this configuration constitutes the use of ATR, at the same time as the length of the fibre has to be kept short (Lourenço et al., 2012). This being said, MIR features greater resolution than NIR spectroscopy, and can therefore quantify components in aqueous solutions at significantly lower concentrations than NIR, in particular when it comes to organic compounds in complex culture media (Landgrebe et al., 2010).

At least in the case of NIR spectroscopy, the applications have evolved from less difficult measurement setups (anaerobic conditions/low agitation) to more complex ones (with intense agitation and aeration), and from at-line or ex-situ to in-situ configurations (Cervera et al. 2009). It seems reasonable to suggest that the scenario has also been similar in the case of MIR spectroscopy. According to Landgrebe et al. (2010), the technical development has reached the point where direct measurement of substrates, products and metabolites, as well as the biomass itself, is generally possible. However, one remaining issue inherently related to the use of IR spectroscopy in bioprocess monitoring is the large absorbance caused by water in the infrared region, which can block out important information present in the spectra (Lourenço et al., 2012). On a more practical and perhaps skill-related level, these systems are also hampered by their own complexity and the complexity of industrial bioprocesses and, in this sense, therefore, they have yet to live up to their full potential (Abu-Absi et al., 2014).

Quantitative measurements of *Pichia pastoris* cultivations using IR spectroscopy have been reported in several publications (e.g., Crowley et al., 2000; Crowley et al., 2005; Guarna et al., 1997; Schenk et al., 2007; Schenk et al., 2008). In general, the conclusion has been that the method is very promising. In this thesis, MIR spectroscopy was applied in order to obtain valuable information about the important substrates glycerol and methanol. The fermentation process studied consisted of three stages using these two substrates. A glycerol batch stage was followed by a glycerol fed-batch stage which, in turn, was followed by a methanol fed-batch stage. The third stage is the production stage, during which in this case the expression of HIV-1 Nef (negative factor) protein took place. This expression is driven by the alcohol oxidase I gene (AOX1), which is strongly repressed in cells grown on glucose and most other carbon sources, but induced over 1000-fold when methanol is used as the only substrate (Cereghino et al., 2002). The glycerol fed-batch should therefore be substrate limited in order to de-repress the AOX promoter in the second stage, whereas care must be taken to avoid accumulation to inhibitory levels of methanol in the third stage, at the same time as complete depletion of methanol is undesirable during the induction phase (Schenk et al., 2007). Thus, the on-line information about the concentration of the substrates is of great interest. MIR spectroscopy could be used to determine the depletion of glycerol at the end of the batch stage, and closed loop control could be used to keep the glycerol concentration at substrate limited levels during the fed-batch stage. Furthermore, and perhaps most importantly, MIR spectroscopy could also be used to keep the methanol concentration at non-inhibitory levels during the third stage. However, this thesis encompasses only the measurements, and no closed loop control was implemented. When it comes to quantitative measurements in *Streptomyces peucetius* fermentations, no other example is found in the literature. However, as a similar application, Roychoudhury et al. (2007) measured glycerol and clavulanic acid at-line in a *Streptomyces clavuligerus* fermentation using MIR spectroscopy.

When it comes to implementation of IR measurements in anaerobic digestion processes, applications of NIR (e.g. Hansson et al., 2002; Hansson et al., 2003; Holm-Nielsen et al., 2007; Holm-Nielsen et al., 2008; Jacobi et al., 2009; Jacobi et al., 2011; Krapf et al., 2013; Lomborg et al., 2009; Madsen et al., 2012; Nordberg et al., 2000; Sarraguça et al., 2009; Ward et al., 2011a; Ward et al., 2011b; Zhang et al., 2009a; Zhang et al., 2009b) seem to be much more common than applications of MIR (e.g. Cuetos et al., 2010; Li et al., 2014; Martínez et al., 2012; Spanjers et al., 2006; Steyer et al., 2002). However, this could be a coincidence rather than an indication that NIR is the favourable technique in AD applications in comparison to MIR. As shown in Chapter 3.3, the ATR technique is very useful for AD measurements; at least in this study, MIR generally performed better than NIR, in particular on samples containing particulate matter.

Many publications describing measurements on the AD process encompass volatile fatty acids (VFA) as constituents of interest. This is because the concentrations of VFAs can indicate organic overload and toxic conditions that inhibit methanogens (Holm-Nielsen et al., 2008). Accumulation of VFAs causes instability in the process, and it is therefore common to construct industrial digesters that are larger than their optimal size (Ward et al., 2011) or, in other words, operate them at an organic load rate far below the optimal one (Jantsch & Mattiasson, 2004). A much more economically favourable solution to the problem of instability would be to monitor and control the system (Jantsch & Mattiasson, 2004). In order to avoid toxic shocks or organic overloads, detection techniques more rapid than those routinely applied today are needed (Hansson et al., 2003); this is probably the main motivation for developing measurement applications for the AD process based on IR spectroscopy.

Instability can occur due to changes in the fermentation environment (Ward et al., 2011) or, perhaps more commonly, due to changes in the actual load rate. In order to keep the process stable, the digester needs to be supplied with an even load of substrate of even quality, or at a rate adjusted according to the quality of the substrate (Jacobi et al., 2011). Agricultural substrates are relatively complex in composition and thereby vary in quality (Jacobi et al., 2011). Thus, there is a need for continuous measurements of the feed and/or the state of the digester. In the context of measurements of the state of the digester, it is the individual rather than the collective VFA concentrations that are of interest (Boe, 2006). Currently, VFAs can be measured using, for instance, straight distillation, steam distillation, a colorimetric technique or gas chromatography, but there is still a need for cheaper and faster methods (Zhang et al., 2009b). This is why this thesis investigates methodologies for faster creation of reliable calibration models for VFAs for both MIR and NIR. Measurement of ammonium was also investigated. This constituent is of interest, since ammonium can directly or indirectly cause inhibition in an anaerobic digestion system (Yenigün & Demirel, 2013).

1.3 Calibration methods and chemometrics

Spectroscopic bioprocess monitoring generates much more data than the information that is actually useful (Lopes et al., 2004). In particular, in-line high frequency measurements generate an enormous amount of data (Biechele et al., 2015). At the same time, the monitoring is also associated with high collinearity in many of the measured constituents and variables (Lourenço et al., 2012). Since it is actually spectra as such that are collected, the spectral data must also be correlated to important process variables (Lourenço et al., 2012). Overall, this creates a need for multivariate data analysis and other chemometrical methods, i.e., mathematical or statistical methods, to analyse data from a chemical system (Lourenço et al., 2012). Chemometrics and multivariate analysis entail their own nomenclature and mathematical methods. However, this nomenclature and these methods are largely considered conventional and well known, and will therefore not be discussed in detail in this thesis.

On a general level, in quantitative measurements using IR spectroscopy, a calibration must be created to describe the relationship between the constituents of interest and the spectra collected from the experimental system. The data used for calibration is commonly (also in this thesis) referred to as the calibration data set. Ideally, this data set will accurately capture the variations and complexity of the system being studied (Abu-Absi et al., 2014). In order to achieve this, three different strategies are commonly used (Cervera et al., 2009): (1) collecting spectra from pure constituent samples or mixtures, (2) collecting spectra from real process samples, and (3) collecting spectra from process samples after the addition of various constituents in order to break correlations. Collecting spectra from pure constituent samples and pure constituent mixtures is usually a fairly quick way of obtaining calibration data, and intercorrelation between constituents can easily be controlled. However, it is unlikely that these types of samples actually capture the variations and complexity of the system being studied. Collecting spectra from actual process samples can potentially overcome this problem; however, generating a sufficient number of samples can be very time-consuming, and the constituent intercorrelation that is likely to occur in bioprocesses cannot be controlled. Collecting spectra of process samples after the addition of various constituents is a method that can potentially capture all the variations and complexity of the process, while at the same time breaking constituent intercorrelation. The method of adding a known amount of the constituent of interest in this way is usually referred to as spiking (Dean, 2003). Although spiking cannot be described as a standard method in quantitative measurements of bioprocesses using IR spectroscopy, it has been used in this context in a number of publications (Finn et al., 2006; Franco et al., 2006; Milligan et al., 2014; Riley et al., 1997; Riley et al., 1998a; Riley et al. 1998b; Roychoudhury et al. 2007; Yeung et al., 1999) and is also used in this thesis.

In addition to the calibration data set, evaluation of a calibration model also requires additional data sets. Basically, three sets are needed in total: One for creating the calibration model, one for testing the model, and one for validating the model (Shaw et al., 1999). The general criterion for the validation data set is that this data set has not been used in any way while establishing the model; this criterion should not be breached unless there is an otherwise insufficient amount of data available (Shaw et al., 1999). In contrast, the test set is typically retrieved from the actual calibration data by means of cross-validation. In leave-one-out cross-validation, one spectrum is removed from the calibration data set, a model is regressed on the remaining data, and a quantitative prediction is made on the removed spectrum. By repeating this procedure, important information on the model's performance is obtained. Note that leave-one-out is only one of many methods used to split the data in cross-validation. Cross-validation is a practical and reliable way to test calibration models, and has become the standard in PLS regression analysis incorporated in almost all commercially-available software (Wold et al., 2001). However, at the same time, cross-validation, particularly when performed as leave-one-out validation, tends to produce overly optimistic results with regard to the model's performance (Kjeldahl & Bro, 2010). Therefore, in order to fully evaluate the model's predictive capabilities, the availability of a separate validation data set is crucial. Thereafter, the model can be used to predict the variables or the process status from on-line (or off-line) spectroscopic data (Biechele et al., 2015).

The mathematical methods used for establishing the relationship between the constituent of interest and the spectral information are called regression methods (Lourenço et al., 2012). As previously mentioned, in bioprocess monitoring using IR spectroscopy, these methods are typically multivariate methods (Biechele et al., 2015). The most common regression methods in this context are principal component regression (PCR) and PLS regression (Lourenço et al., 2012). However, the use of multiple linear regression (MLR) has also been quite common. Although the popular misconception that MLR is only possible for one dependent variable is untrue, singularity is still a frequent problem when using MLR (Geladi & Kowalski, 1986). However, MLR is still the simplest form of performing an inverse multivariate calibration (Næs et al., 2002). Unlike MLR, PLS can analyse strongly collinear and noisy data, with numerous independent variables (Wold et al., 2001). Another desirable feature of PLS is that the precision of the model parameters improves with the increasing number of relevant variables and observations (Wold et al., 2001). Just like PLS, PCR is also a factor-based method capable of being a full-spectrum method. Both PCR and PLS compress the data into linear combinations of the original spectral data. These linear combinations are referred to as factors, PLS components, or loading vectors in the case of PLS, and almost always as principal components (PCs) in the case of PCR and principal component analysis (PCA). The main difference between the methods is that PCR models the spectra without using constituent information, whereas PLS maximises the covariance between the spectral information and the constituent (Thomas & Haaland, 1990). As a consequence, PCR often needs more components than PLS in order to achieve the same accuracy.

A common issue when applying regression methods is how to select the appropriate wavelengths or wavelength intervals, a process often referred to as variable selection. This is because there are wavelengths in the spectra unlikely to contain any relevant information, at the same time as the peaks may be neither distinct nor sharp (Soons et al., 2008). There are typically numerous variables at hand, simply because the instrument provides them, resulting in a situation in which the variable selection can have a very significant impact on model performance (Kjeldahl & Bro, 2010). This being said, variable selection has received little attention in the context of bioprocess monitoring using IR spectroscopy, and this is true also for this thesis. Commonly, yet rarely explicitly stated, variable selection takes place by discharging the noisy parts of the spectra, i.e., the end intervals and regions with very high absorbance. A slightly better method might be to base the variable selection according to where the pure constituent appears in the spectra (e.g. Vaidyanathan et al., 1999). However, this methodology may be hampered by the fact that constituents in a complex mixture may contribute to signals that are spread across the complete spectral range (Soons et al., 2008). More advanced variable selection methods certainly do exist, e.g. Brink & Westerlund (1995), but such advanced variable selection methods are not a part of this thesis.

As previously mentioned, this thesis encompasses the use of local calibration methods in the context of AD monitoring. In NIR spectroscopy, many studies have been carried out using local calibration methods. However, in contrast, no examples have been found in which local models were used in connection with MIR spectroscopy. Even within the field of NIR spectroscopy, there is no distinct universal definition of a local calibration model. This being said, in most cases the term corresponds with the definition that "a local model is one built using only local features" as stated by Kolari et al. (2006). This is also the definition of a local model used in this thesis. In practice, a local model could also be defined as a model that has been regressed on less than all the available calibration data. However, in order to fully comply with the "only local features" requirement, the data used for the regression of the local model cannot be selected randomly from all the available calibration data. Some type of similarity indices must be used in order to identify the part of the calibration data containing the local features. These similarity indices are mostly referred to as distance measures in this thesis. It also seems reasonable to suggest that the main differences among the established local calibration methods are found in which distance measure is used in the selection of the local subset of calibration data. Distance measures in spectroscopy can, for instance, be expressed by means of correlation, absolute deviation, Euclidean distance or Mahalanobis distance. Acknowledging some exceptions, these distance measures can in turn be applied in, for instance, spectral space as well as PC and PLS score space. Thus, with the distances already mentioned, there are numerous ways to establish similarity indices in IR spectroscopy.

In quantitative NIR spectroscopy, new calibration methods are constantly developed and evaluated. Therefore, many local calibration methods have also been published. However, only a few of the more well-known methods will be mentioned in this thesis. The first local calibration method used in the context of NIR spectroscopy was probably CARNAC (Davies et al., 1988). This method uses the square of the correlation as a distance measure or similarity indices. The square of the correlation is calculated on weighted Fourier coefficients of the spectra. As such, the actual prediction in CARNAC is not a multivariate method. The quantitative prediction is instead calculated as a similarity-weighted average of the constituent information assigned to the spectra from the selected local subset of samples.

One of the older and more well-known local calibration methods is found in the form of LOCAL (Shenk et al., 1997). This method is available in commercial calibration software, and has been used in many studies. In this method, spectral correlation serves as the distance measure and a local PLS model is regressed on the selected spectra. Although not studied in any greater extent, but with a sufficient number of citations, locally biased regression (Fearn and Davies, 2003) can also be seen as one of the established local calibration methods. Locally biased regression differs from the two local methods previously mentioned by, among other things, applying two criteria in the selection of the local data subset. One criterion states that the score on the first orthogonal signal correction (OSC) factor must be inside a specified interval compared to the unknown spectrum. The other criterion is a distance measure based on the distance to prediction made by a global PLS model. The method of obtaining the prediction also appears to be unique; this is probably the only local calibration method found (at least in terms of NIR spectroscopy) that does not make the prediction directly based on information extracted from the local data set. Instead, the bias and skew for the global PLS model on the local subset is determined. The prediction is thereafter obtained as either a bias or skew and bias corrected prediction made by the global model.

At least in terms of number of citations of the original publication of the method (Næs et al., 1990), LWR seems to be the most well-known local calibration technique within NIR spectroscopy. Although Næs et al. (1990) introduced this method to the field of NIR spectroscopy, its origin is found elsewhere. Cleveland (1979) first published this methodology as a concept for smoothing scatterplot, and explained the functionality with "A robust fitting procedure is used that guards against deviant points distorting the smoothed points". This publication by Cleveland (1979) included applications quite far from NIR spectroscopy, for instance, lead intoxication and abrasion loss. LWR is used in Publication IV in this thesis, and is of particular interest in the sense that it is similar to the multi-layer PLS method also addressed in this thesis. Næs et al. (1990) used the weight function suggested by Cleveland and Devlin (1988), and the Mahalanobis distance in PC space for the selection of the local subset. It was concluded that LWR performed significantly better than PCR in two out of three examples studied. This is perhaps what inspired further development or modification of the method by, for instance, taking into account the prediction capability of the PCs when selecting the local subset (Næs and Isaksson, 1992); including the dependent variable as a distance measure (Wang et al., 1994); performing PCR directly on the local data (Aastveit and Marum, 1993); and by using PLS regression for obtaining the local calibration model (Shenk et al., 1997). LWR as implemented in the PLS Toolbox (Eigenvector Research, Inc. 3905 West Eaglerock Drive, Wenatchee, WA 98801, USA) for MATLAB (The MathWorks AB, Kista, Sweden) features two distance measures. One is the auto-scaled distance in PC space and the other the Euclidean distance to prediction. The relative weight of these distances can be freely chosen; thus, the selection can be based on either one of these two distances or a weighted combination of them. The model can be regressed by means of PLS, PCR or on global PCs. The algorithm also allows for reselection of the local data set based on the prediction of the local model. However, in contrast to the ML-PLS method, it is always the same number of local samples that are selected.

This thesis also includes the creation and further refinement of the ML-PLS calibration method, which is a local calibration method. The key and novel feature of ML-PLS is an iterative or stepwise approach during which the local calibration data set is sequentially reduced in size. The name of the method derives from the naming of each step or iteration as a layer. Although the current version of ML-PLS, as described in Chapter 3.4 and in Publication V, is in fact an iterative method, the original concept featured a fixed set of models organised in a layer structure. Thus, at this point the term layer as such is not very descriptive, but perhaps rather a reminder of the origin of the concept. The terminology used in connection with ML-PLS also includes the terms design, structure, and total model. In this context, ML-PLS structure and design actually mean the same thing, which is how the calibration data set should be reduced from layer to layer and how many layers are included in the model. The term total model, in turn, describes the whole design, and the output of the total model is the prediction made on the last layer in the design. Later development of ML-PLS also focuses on the usability of different distance measures.

Chapter 2

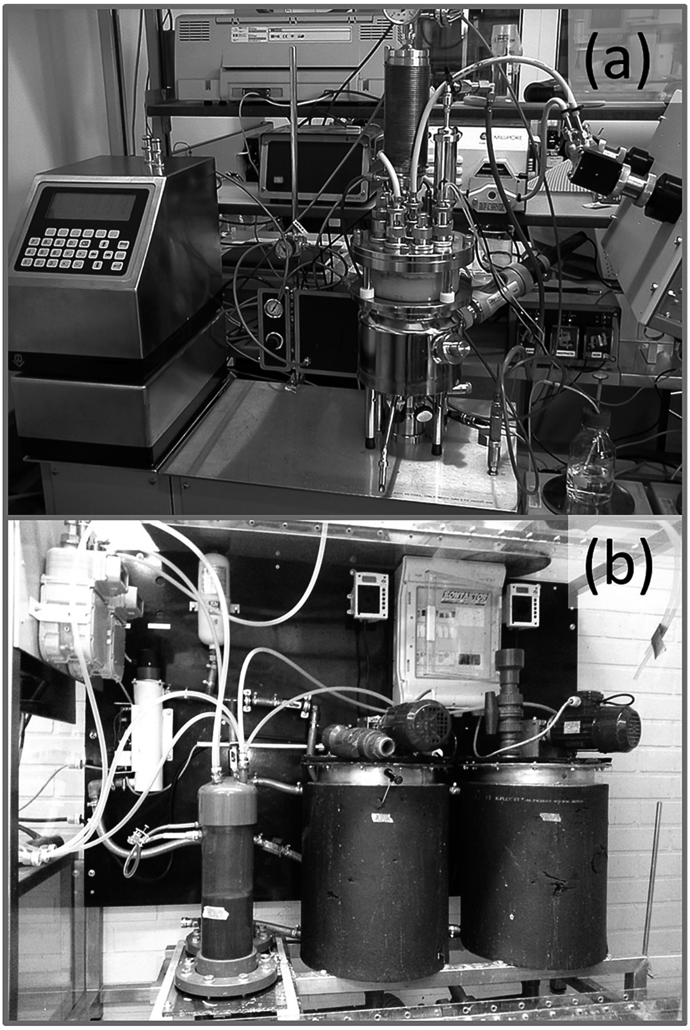
2. Materials and methods

2.1 Fermentation processes

The *Pichia pastoris* fermentations were carried out in a 3.5 L Chemap CF 3000 fermenter, and the content volume was 2 L after inoculation. The volume of the inoculum was 0.2 L, and the medium volume 1.8 L. The fermentation was performed in three separate stages; a glycerol batch phase was followed with a glycerol fed-batch, and the induction phase was a methanol fed-batch fermentation. The pH and the dissolved oxygen tension were kept at a constant level throughout the fermentation. Cultivations were carried out with both X33 (phenotype Mut^{+}) and KM71H (phenotype Mut^{S}), and they were performed for production of recombinant HIV-1 Nef protein (Sirén et al., 2006). The fermentation methodology is described in detail in Vermasvuori et al. (2009). In the Mut^{S} cultivation, the temperature was kept constant at 30 °C, whereas in the Mut^{+} cultivations the temperature was decreased during the production phase in order to slow down the proteolysis of the product.

The batch fermentations of *Streptomyces peucetius* var. *caesius* N47 were carried out in the same fermenter. The content volume was 2 L after inoculation. The inoculum volume was 5% of the total volume. During the fermentations, the temperature was kept at 30 °C, the aeration at 1 VVM (volume of air to volume of medium per minute), the agitation speed at 300 rpm (revolutions per minute), and the absolute pressure at 1.5 bars. The pH was kept constant at 7 by adding a 4 M NaOH solution. Two fermentations were carried out and two different medium compositions were used. One contained soy peptone and is therefore referred to as the SP medium. The other contained Pharmamedia and is therefore called the PM medium. The main constituents in both media were glucose and starch. The inoculum was cultivated on the SP medium for both fermentations.

The anaerobic digestions were carried out in two custom-built continuously stirred 38 L laboratory scale reactors, with a working volume of 27 L. These thermophilic anaerobic digestions were started with a culture obtained from a municipal waste treatment plant (Ab Stormossen Oy), and the inoculum volume was 3 L. The substrate was a 4% dry weight mixture of pig manure, waste from industrial treatment of raw fish, and greenhouse plant waste, at dry weight ratios of 2/3, 1/6, and 1/6. This thesis encompasses samples from four separate digestions. Picture 2.1 shows the two reactor systems used in this thesis.



Picture 2.1. Reactors used in (a) the fermentations of *Pichia pastoris* and *Streptomyces peucetius*, as well as (b) the anaerobic digestions.

2.2 Reference measurements

The reference measurements of the glycerol, methanol, and glucose concentrations were done with a high-performance liquid chromatograph (HPLC) from Waters (Millipore Corporation, USA) equipped with a Guard-Pak Pre-column and a FAM-Pak column. The samples were centrifuged and the supernatant filtered before injection. Either the reference measurements were performed directly after sample removal from the reactor, or the filtrate was stored in a freezer for later determination. The area of the chromatogram peak for each component was converted into concentration readings. At later stages in the *Streptomyces peucetius* fermentations, the glucose concentration could not be determined with the HPLC due to overlapping peaks. Instead, an Accutrend® blood sugar sensor (Buehringer Mannheim Scandinavia AB, ref. 1544 179/680), and an YSI 2700 D Select (Yellow Springs Instrument, Yellow Springs, Ohio) enzymatic membrane analyser were used. In the anaerobic digestions, the ammonium concentration was measured using flow injection analysis and photometrical detection (FIAstar 5000 Analyzer, Foss Tecator, Denmark), by the gas permeable membrane method in accordance with the EN ISO 11732:2005 procedure. The volatile fatty acid concentrations were measured using a gas chromatograph mass spectrometer (Shimadzu QP-2010 GC/MS, Shimadzu Scientific Instruments, 7102 Riverwood Drive, Columbia, MD 21046, USA). The TVFA concentration was computed as the ion concentration sum of acetic acid, propionic acid, isobutyric acid, butyric acid, 4-methyl valeric acid, valeric acid, 3-methyl valeric acid and hexanoic acid.

2.3 Spectroscopic measurements

In the collection of the MIR spectra, an ASI ReactIR 1000 Fourier transform infrared spectrometer (ASI Applied Systems, Millersville, MD) was used. The ReactIR was equipped with a liquid nitrogen cooled mercury cadmium telluride (MCT) mid band detector with a 12-hour hold time. The ATR was a DiComp (diamond composite) probe, adapted to standard Ingold fittings and mounted on a 42-inch articulated arm. The NIR measurements were carried out with a portable HandySpec Field diode array instrument (tec5 AG, In der Au 27, 61440 Oberursel, Germany). The instrument was equipped with an MMS1 (monolithic miniature spectrometer) detector for the lower wavelengths and a PGS2.2 (plane grating spectrometer) detector for wavelengths above 1000 nm, comprising 256 sensors in the region 305-2200 nm. The AgroSpec software (tec5 AG, In der Au 27, 61440 Oberursel, Germany) was used as an interface in the collection of the spectra. An in-house-built 5-mm flow-through transmission cell was plugged in one end and used as a cuvette. Picture 2.2 shows the MIR and NIR spectrometer when set up for off-line measurements of anaerobic digestion samples.

Picture 2.2. The (a) ReactIR and the (b) tec5 instrument when set up for off-line AD monitoring.

2.4 Mathematical manipulation of calibration spectra

The construction of the calibration data set for the methanol calibration model used in the *Pichia pastoris* application relied on two fermentation spectra collected when the concentration of methanol in the reactor was determined to be zero. One of these spectra was collected before the beginning of the methanol fed-batch phase, and one at the end of the methanol fed-batch phase during a fermentation in which the temperature was lowered during the production phase. These spectra were considered to represent the features of the absorbance spectra that are present during the methanol fed-batch phase, but that are unrelated to the methanol concentration. These spectra should then incorporate features from the dilution of the reactor content due to the addition of the substrate, the decrease in temperature, and the secretion of new compounds into the liquid phase. In order to increase the spectral feature's span of the background matrix, a number of calculations were performed on the two spectra. A total of nine new spectra were defined or computed.

The 9 spectra of the simulated background matrix ($SS_{1\text{-}9}$) were calculated from the two fermentation spectra (FS_1 and FS_2) as follows:

$$SS_1 = 0.75\ FS_1$$
$$SS_2 = FS_1$$
$$SS_3 = 1.25\ FS_1$$
$$SS_4 = 0.75\ FS_2$$
$$SS_5 = FS_2$$
$$SS_6 = 1.25\ FS_2$$
$$SS_7 = \frac{SS_1 + SS_6}{2}$$
$$SS_8 = \frac{SS_2 + SS_5}{2}$$
$$SS_9 = \frac{SS_3 + SS_4}{2}$$

All pure methanol spectra, from 10 pure methanol samples, was then added to each of the simulated background matrix spectra, thus yielding 90 spectra for the training data set. This strategy essentially mimics the way the methanol features appear in real fermentation spectra.

The mathematically manipulated calibration data set used for glycerol measurements was constructed by computing a minimum correlated scheme for the interfering compounds (sulphate, phosphate, and ammonia) and glycerol. Since the addition of the spectra was carried out "manually" and one spectrum at a time, the scheme was limited to a total number of 40 combinations of the 4 compounds at 5 concentration levels. The concentration span for the interfering compounds roughly represented the concentrations that were present during the cultivations. The absorbance spectra of the concentrations corresponding to the scheme were then added together. In addition to the pure glycerol spectra, these 40 synthetic multi-constituent spectra were then included in the training data for the glycerol calibration.

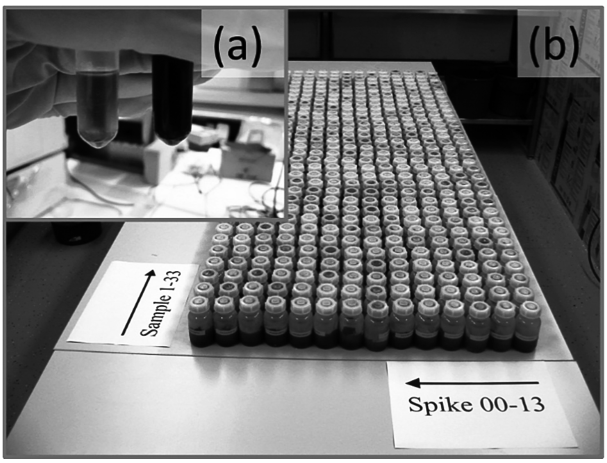
The calibration data set for glucose, acetate and starch in the *Streptomyces peucetius* fermentations was obtained as follows. Spectra were collected at 5 equidistant concentrations for each of the constituents glucose, starch and sodium acetate, and at 3 equidistant concentrations for each of the constituents pharmamedia, soy peptone, yeast extract, $MgSO_4$ and KH_2PO_4. From these pure component spectra, synthetic multi-constituent spectra were computed according to a full factorial design with 6 levels for glucose, starch and acetate, and 3 levels for pharmamedia, $MgSO_4$ and KH_2PO_4 for the PM medium, and for the SP medium by replacing the pharmamedia with soy peptone. The first level for glucose, starch and sodium acetate in the design represented 0 concentration and absorbance. The remaining levels and the three levels for pharmamedia, soy peptone, $MgSO_4$, and KH_2PO_4, were represented by concentration and spectral data of the constituent in question. In this way, two data sets (one for the PM medium and one for the SP medium) with 5832 spectra, and corresponding concentration information in each, were obtained.

2.5 Spiking scheme for anaerobic digestion samples

The spiking of the AD process samples with ammonium, acetate and propionate was carried out as a three-factor central composite design, with axial points at the cube walls and no centre points. The length of the cube wall was 2 in coded space, equivalent to 5 g L^{-1} in concentration space when added to a sample with zero constituent concentration. The design itself is fully orthogonal, but becomes somewhat skewed when implemented. This occurs because the process samples always contain constituents at some concentration, and this initial concentration will become diluted when spiking is performed. Table 2.1 shows the design represented as coded values and spike size in concentration space. The process sample, as extracted from the reactor, represents the coordinate (-1,-1, -1) in the coded design. The 33 process samples, each one split into 14 subsamples, amounted to 462 samples for IR spectroscopic measurements. Picture 2.3 shows the impact of centrifugation on the AD samples, and the full series of spiked samples.

Table 2.1. The spiking design used in the AD measurement application given as coded values and spike size in concentration space.

Coded coordinates			Spike size [g L^{-1}]		
Ammonium	Acetate	Propionate	Ammonium	Acetate	Propionate
-1	-1	-1	0.0	0.0	0.0
-1	-1	1	0.0	0.0	5.0
-1	1	-1	0.0	5.0	0.0
-1	1	1	0.0	5.0	5.0
1	-1	-1	5.0	0.0	0.0
1	-1	1	5.0	0.0	5.0
1	1	-1	5.0	5.0	0.0
1	1	1	5.0	5.0	5.0
0	0	-1	2.5	2.5	0.0
0	0	1	2.5	2.5	5.0
0	-1	0	2.5	0.0	2.5
0	1	0	2.5	5.0	2.5
-1	0	0	0.0	2.5	2.5
1	0	0	5.0	2.5	2.5



Picture 2.3. The (a) impact of centrifugation on the AD samples and (b) the full set of spiked AD samples.

2.6 Nomenclature in multivariate modelling

In this thesis, the performance of the quantitative models is typically given in the form of root mean square error of calibration (*RMSEC*), root mean square error of cross-validation (*RMSECV*), and root mean square error of prediction (*RMSEP*). However, Publication I uses the standard error of prediction (*SEP*) instead of *RMSEP*. The mathematical definitions of these can be found in Næs et al., (2002). Qualitatively, *RMSEC*, *RMSECV* and *RMSEP* are estimates of the prediction error obtained on the calibration data, in cross-validation, and on the validation data respectively. The difference between *RMSEP* and *SEP* is essentially that *RMSEP* also includes the bias. Publication III also uses the ratio of standard error of performance to standard deviation (*RPD*) and the range error ratio (*RER*). Qualitatively these are relative performance indicators or errors, for *RPD* relative to the standard deviation and for *RER* relative to the measured constituent span. The mathematical definitions of these can be found in Fearn (2002). Standard mathematical manipulation of the spectral and constituent data is always applied in this thesis before modelling. These methods are called pre-processing or pre-treatment. The methods used in this thesis are mean centring, auto scaling, two-point baseline correction, smoothing and derivatives. All smoothing and derivatives are computed using the Savitzky–Golay method (Gorry 1990), in which smoothing is achieved by fitting successive subsets of adjacent data points with a low-degree polynomial. When used to compute derivatives, the fitted values are used instead of the actual spectral values. Derivatives are useful, for instance, in baseline removal. Taking the first derivative removes an additive baseline, and taking a second derivative removes a linear baseline (Næs et al., 2002).

2.7 Methodology used in the comparison between NIR and MIR

In order to compare the usability of NIR and MIR spectroscopy in AD process monitoring, spectra were collected using both the NIR and the MIR instrument on samples before and after sample centrifugation. Two spectra from each sample were used in the comparison. The samples that had not been centrifuged are hereafter referred to as the unprepared samples. In the NIR measurements on the unprepared samples, the quality of the spectra was very dependent on how long the samples had been allowed to settle in the sample bottle or in the cuvette; as a result, the reproducibility was very poor. Therefore, from the 924 spectra collected in total, 131 were removed as outliers. This decision was based on abnormally large prediction residuals when using the robust PLS algorithm in the PLS toolbox. Only after this outlier removal, the data were split into model regression and validation data sets.

For each of the constituents acetate, ammonium, propionate and TVFA, the spectral and constituent data were sorted in descending concentration order. Starting from the second sample, every third sample (i.e., two spectra) was thereafter selected as validation data and the remaining samples as model regression data. For each of the constituents, a similar split was also made for a concentration range covering only the lowest to the highest concentration in the actual process samples with an extrapolation of 10%. Thus, for each of the two instruments and each of the two sample preparation methods, model regression and validation data were obtained for the natural concentration range in the process samples as well as the full concentration range of the spiked samples (i.e., 64 sets of spectral data). The data split is visualised in Figure 2.1.

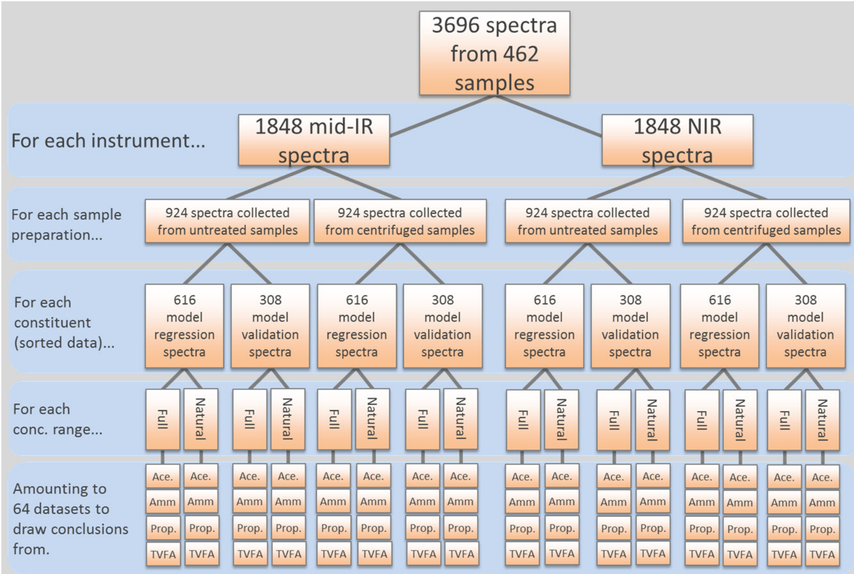


Figure 2.1. Data sets used in the comparison between NIR and MIR for anaerobic digestion monitoring.

In order to reduce the impact of human decisions on the conclusions from this study, all model parameters for all data sets were determined using the same steps, and the decisions were based on results from cross-validations. These steps were (1) wavelength interval selection, (2) pre-processing selection, (3) regression data outlier removal, and (4) new pre-processing selection. For each data set, the model with the lowest *RMSECV* was thereafter validated, including a step of validation data outlier removal. This whole process was carried out by means of MATLAB scripts utilising the PLS toolbox. The wavelength region selection was carried out by first selecting a narrow spectral interval of primary interest; moving one of the interval ends at a time towards the end of the spectra; computing *RMSECV*-values for every second data point; reselecting the end points according to the lowest *RMSECV* value; and repeating the whole procedure three times with either mean centring or auto scaling of the spectra.

The best pre-processing alternative was evaluated by studying the *RMSECV* values for models regressed using the following spectral smoothing or derivative options:

- neither derivative nor smoothing

- Savitzky-Golay smoothing
- first order Savitzky-Golay derivative
- second order Savitzky-Golay derivative

The smoothing or derivatives were based on second order polynomials, with a number of points from 3 to 63 at a step size of 2. Each of the options described above was then followed by either mean centring or auto scaling (138 settings in total). Based on the predictions of the regression data with the model using the spectral pre-processing with the lowest *RMSECV* value, 10% of the regression data was then removed as "outliers" according to the size of the residuals. Thereafter, all 138 pre-processing settings were re-evaluated, and the pre-processing setting with the lowest *RMSECV* value was selected as the pre-processing setting in the model used for the predictions of the validation data set. Based on the predictions made with this model on the validation data, 10% of the validation data was removed as "outliers" according to the absolute size of the prediction residuals. For clarification, removing 10% of the spectra as "outliers" was not based on an assumption that 10% of the spectra were outliers; it was based on the assumption that removing 10% of the data should in effect remove any outliers in the data.

Chapter 3

3. Results and discussion

Most of the information provided in this chapter can also be found in Publications I-V. In fact, many things are presented in greater detail in these publications. However, this chapter also contains information in addition to that found in these publications. For instance, the comparison between NIR and MIR spectroscopy in the context of AD monitoring is not included in any of the publications. In addition, Publications IV and V describe the use of ML-PLS, but not how the method has evolved. In any case, the core of this chapter is based on Publications I-V, and describes the usability of IR spectroscopy in the context of quantitative bioprocess monitoring.

As a form of recap from Chapter 1, and as a short summary of the publications included in this thesis; Publication I describes on-line measurements of the glycerol and methanol concentration in *Pichia pastoris* fermentations; Publication II on-line measurements of the glucose, acetate and starch concentrations in *Streptomyces peucetiu*s fermentation; Publication III off-line measurements of the ammonium, acetate, propionate and TVFA concentrations in AD; Publication IV the usability of local calibration techniques for measurement of the ammonium and acetate concentrations in AD; and Publication V how well the latest version of ML-PLS performed in terms of measurement of the ammonium concentration in AD. Since the methodology used to implement these measurement applications is in many cases somewhat unusual, this aspect receives a fair amount of attention. However, in the end, it is not necessarily the methodology used that is of greatest interest, but rather the implications of successful applications of bioprocess monitoring using IR spectroscopy for the future development of the biotechnology sector.

3.1 Mathematical generation of calibration data

The potential of using simple spectral additions to create calibration data superior to using only pure component samples was evaluated in the form of on-line measurements in fermentations of both *Pichia pastoris* and *Streptomyces peucetius*. As mentioned in Chapter 1, the ability to obtain on-line information about the substrate concentrations in *Pichia pastoris* fermentations could potentially be of great importance for process development and efficient production. In the case of the batch fermentations of *Streptomyces peucetius*, the benefit of on-line information about the substrate concentrations may, however, be limited to increased process knowledge. Nevertheless, increased process knowledge is always beneficial and potentially very valuable, although the types of benefits and value may not be obvious at this point.

The spectral variance caused by a single spectrum can in PLS regression be explained to 100% by a single PLS component. In this sense, spectral addition may seem redundant. However, PLS models the quantitative constituent of interest simultaneously with the spectral information, so the concentration information will always contribute to the formulation of PLS components. In this sense, spectral addition can be of interest. A trivial, yet important motivation for using spectral addition when doing model regression is the fact that selection of model parameters, such as wavelength selection and base-line points, many times takes place by means of visual inspection of the spectra by a (skilled) instrument operator. Thus, having spectra in the calibration set that mimic spectra from the actual process is also of significant importance in itself.

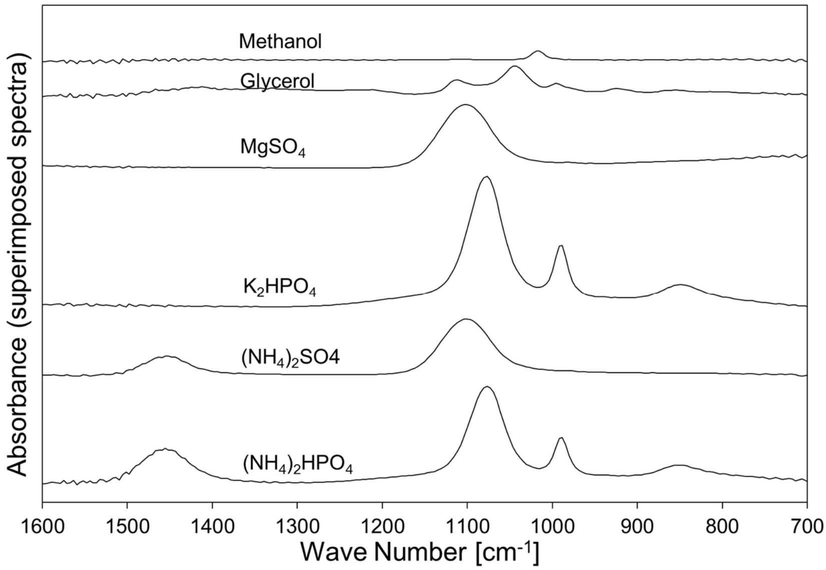


Figure 3.1. Mid-infrared absorbance spectra (superimposed) of the substrates methanol and glycerol, and the other main constituents of the medium used in the *Pichia pastoris* fermentations.

Figure 3.1 shows the absorbance spectra of the main constituents in the *Pichia pastoris* fermentations at relevant concentrations. The situation can be described as typical in bioprocess monitoring with MIR spectroscopy. Distinct features can be seen for each constituent, but the information is overlapped in the sense that several constituents cause absorbance in the same region of wave numbers. At the same time, the features from the constituents of interest (glycerol and methanol) make up only a fraction of all the features of the fermentation broth in terms of absorbance. In the case of glycerol measurement in the *Pichia pastoris* fermentations, most of the spectral variance that occurred during the glycerol batch phases could be attributed to changes in the glycerol concentration. Therefore, a PLS model regressed on pure glycerol spectra actually performed rather well, although not as well as the model based on the synthetic multi-component spectra. The impact of the additional information obtained with the synthetic multi-component spectra is more clearly visible in the PLS components of the respective model, i.e., the model based on only pure component spectra and the model based on the synthetic multi-component spectra. Figure 3.2 shows the three first components of these two models. Clearly, the additional information affects both the information in the PLS components and the noise level. The model based on the synthetic multi-component spectra performed well in several fermentations (four are documented in publication I), and the *SEP* was determined to be 0.7 g L^{-1}.

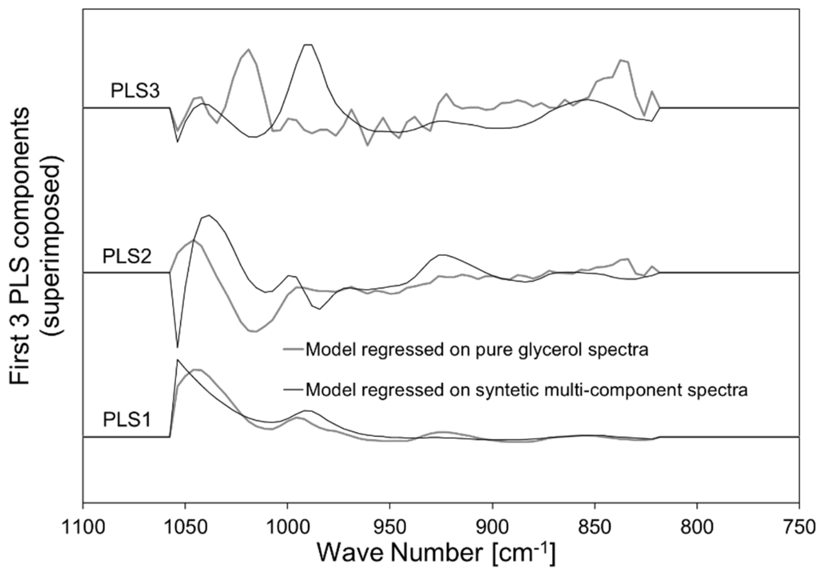


Figure 3.2. Comparison between the three first PLS components for two glycerol PLS models, one regressed on pure glycerol spectra only and the other regressed on synthetic multi-component spectra.

In the case of measurement of the methanol concentration in the *Pichia pastoris* fermentations, no comparison was made between the model based on the mathematically manipulated spectra and a model based on pure component spectra. Since only a tiny fraction of the spectral absorbance in the wavelength region at which methanol absorbance takes place could be attributed to methanol absorbance during the fermentations, the only calibration made was regressed on the mathematically manipulated spectra, visually utilising the calculated spectral features when establishing the end points for a baseline used in the model. Another issue was the fast consumption of methanol when using Mut^+, and therefore the actual model validation was performed on a Mut^S fermentation. Although validation was not possible with Mut^+, Publication I shows that the predicted concentrations could be explained by changes in the feed rate. Figure 3.3 also appears to reveal some interesting information that could hardly have been obtained without on-line measurements. The figure shows that methanol momentarily starts to accumulate at the beginning of the methanol fed-batch. This seems reasonable, since the cells have to adapt to a new substrate. Furthermore, methanol seems to momentarily accumulate three more times during this fermentation. As it turned out, these occasions coincide with sample withdrawal. This observation also makes sense, since the method used for sample withdrawal effectively reduced the head pressure to zero and thereby limited the concentration of dissolved oxygen. Thus, lack of oxygen during sampling implies accumulation of methanol; this observation could not have been made without on-line measurements. Even if a sample could have been withdrawn at the "right time", the methanol in the sample would probably have been depleted through ongoing microbial activity by the time the sample was injected into the HPLC.

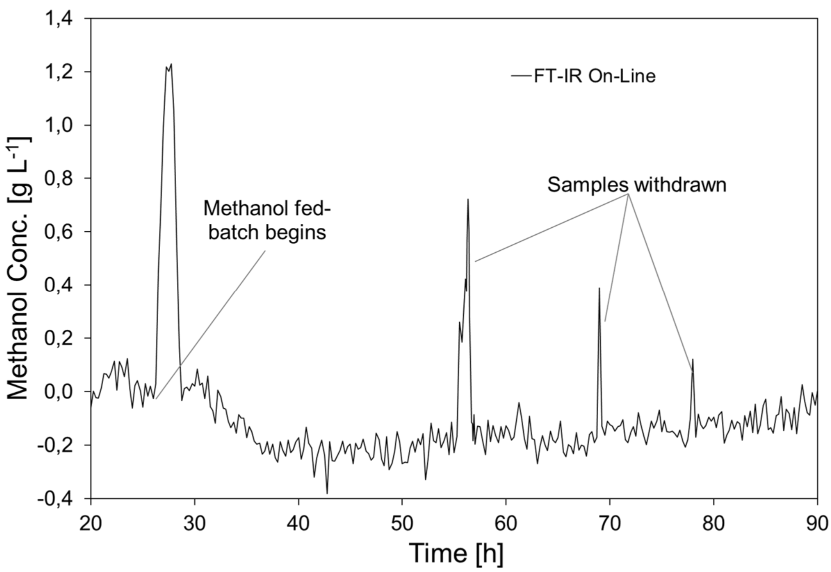


Figure 3.3. MIR methanol predictions in a *Pichia pastoris* phenotype Mut^+ fermentation.

The on-line predictions of methanol performed by MIR in a Mut^S fermentation are shown in Figure 3.4. *The SEP* was determined to be 0.13 g L^{-1}. However, Figure 3.4 is perhaps more informative than the actual *SEP* value itself, in the sense that the remarkable accuracy of the on-line predictions is obvious. Thus, this is without a doubt an application in which MIR measurement can readily be applied and thereby significantly increase the potential for process improvement.

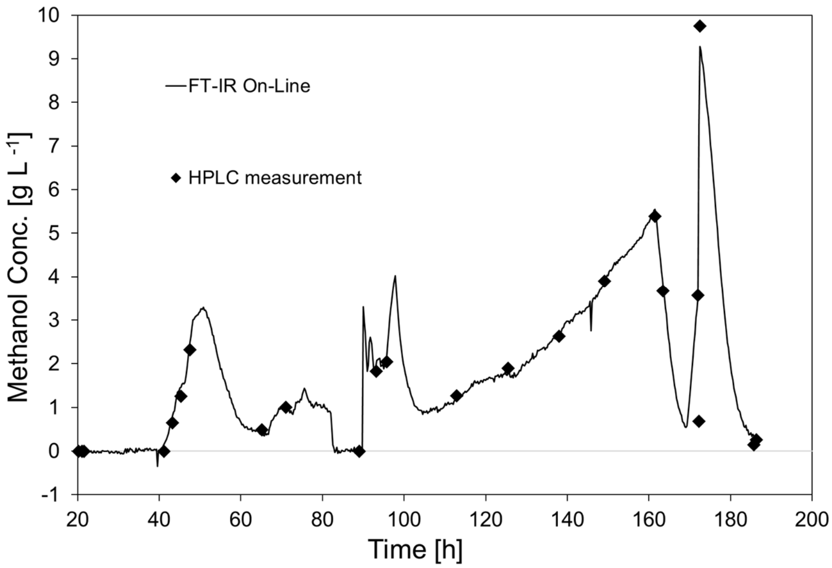


Figure 3.4. MIR methanol predictions in a *Pichia pastoris* phenotype Mut^S fermentation.

Whereas the *Pichia pastoris* fermentations were performed as three separate stages, the *Streptomyces peucetius* fermentations were carried out as "simple" batch fermentations. Thus, the consumption of the substrates (glucose and starch) will be correlated to each other, and, in turn, also negatively correlated to metabolic products (such as acetate). Other issues associated with implementing the quantitative measurement of glucose, starch and acetate were the spectral similarity of starch and glucose, the fact that starch does not have a readily definable molecule (being a polymeric carbohydrate consisting of a large, but unspecified, number of glucose units joined by glycosidic bonds), and the low concentration and thereby low absorbance of acetate. Furthermore, starch becomes significantly more soluble in water when the water is heated, which in practise means that the spectral features change during the heat sterilisation process.

Figure 3.5 shows the first PC retrieved with PCA from the actual process spectra sub-plotted with spectra of the constituents of interest, i.e., glucose, starch and acetate. Apparently, the first component contains a combination of information from glucose and starch, and no acetate information. Furthermore, the features of starch and glucose are completely overlapping, although distinct features also can be found. This being the case, a calibration based on real process spectra as such would probably not work for this application. In Figure 3.5, the acetate concentration is at 2 g L^{-1}, roughly the highest concentration obtained during the fermentations. As can be seen, the absorbance is regardless very low. Consequently, this preliminary observation already suggests that measurements of the acetate concentration would be challenging to implement.

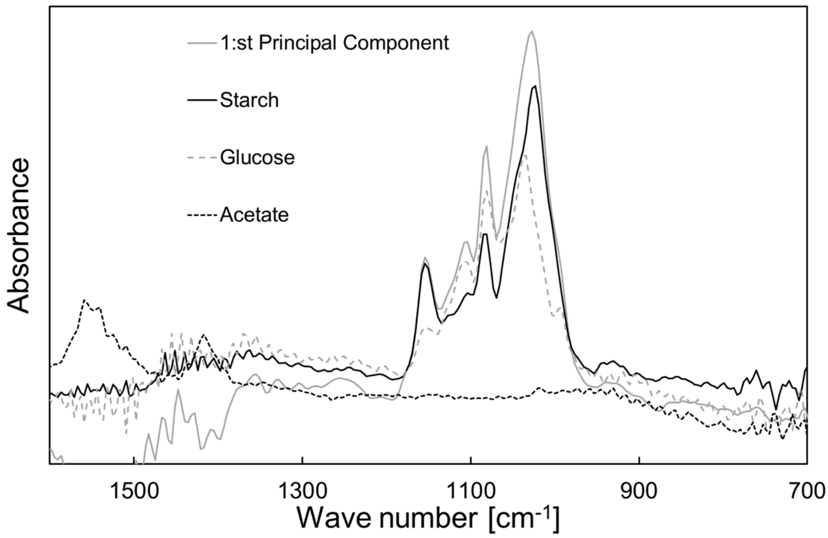


Figure 3.5. The first principal component (rescaled) retrieved from the validation data from the *Streptomyces peucetius* fermentation with the PM medium, plotted together with a spectrum of glucose (10 g L^{-1}), starch (10 g L^{-1}), and sodium acetate (2 g L^{-1}).

Since most of the spectral changes that occurred during the initial consumption of glucose could be attributed to the change in glucose concentration, the model regressed on pure component spectra performed very well in the PM medium fermentation, actually almost as well as the model regressed on the calculated spectra (with *RMSEPs* of 1.5 and 1.3 g L^{-1} respectively). However, in the SP medium fermentation, where the absorbance increased by approximately 300% during heat sterilisation, the model regressed on pure component spectra performed considerably less well than the model regressed on calculated spectra (with *RMSEPs* of 2.5 and 1.7 g L^{-1} respectively). Thus, the impact of the additional calibration data is more substantial when the spectral features start to diverge more from what is found in a calibration data set containing only pure component spectra.

In the case of implementing acetate measurements, some extraordinary methods were needed to establish a somewhat reliable model. Here an *RMSEP* of 0.2 g L^{-1} was obtained for the PM medium using the model based on calculated spectra, whereas models based on pure constituent spectra were essentially useless. Figure 3.6 shows the on-line measurement results with the PM medium fermentation using the models based on calculated spectra. An interesting observation is that both the MIR and the Accutrend reference measurements start to predict an increase in the glucose concentration after about 50 h. However, the more reliable YSI meter shows that the glucose concentration is actually zero, as it should be. Thus, the on-line measurement is tracking the concentration of some unknown intermediate between starch and glucose, which certainly can be seen as increased process knowledge.

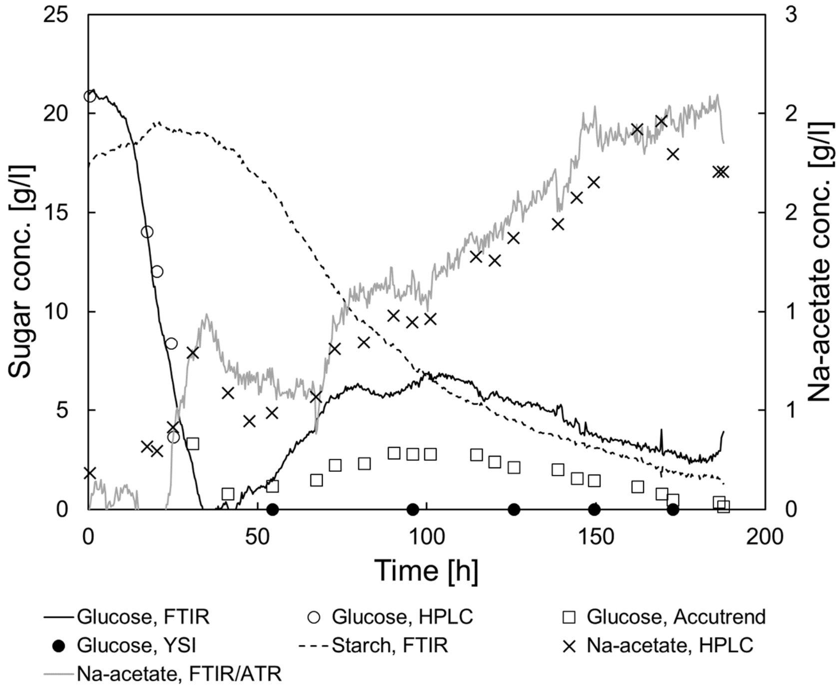


Figure 3.6. On-line measurement of the glucose, starch and sodium acetate concentrations in the fermentation with the PM medium.

3.2 Pure component spiking of process samples

According to Landgrebe et al. (2010), evaluation based on a linear combination of spectra obtained for each individual component is not possible in many bioprocess applications. Furthermore, spectral addition to create a calibration data set comes with other limitations. The variance of a given number of spectra will be explained to 100% by the same number of PLS components; therefore, actual chemical spiking of samples has its benefits. However, this also comes with its own risks in the form of potentially inducing chemical reactions in the sample. It seems reasonable to suggest that in bioprocesses, where the medium can already consist of more than 50 constituents and a number of known and unknown metabolites, the benefit of adding a chemical spike to the sample is somewhat uncertain. However, when successful, a chemical spike should generate spectral information beyond that of spectral addition. Based on these conclusions, the spiking methodology study presented in Publication III was therefore initiated with the pragmatic approach that this methodology would either work or not.

The use of a spiking scheme on the AD samples was inspired by a correlation matrix shown in Holm-Nielsen et al. (2007). The matrix contained the correlation between 11 constituents in samples collected from full-scale anaerobic digestion methane production. A total of 13 intercorrelations were above 0.8. Thus, it was assumed that constituent intercorrelation would be a severe problem when implementing quantitative measurements based on NIR spectroscopy. At the time, the usability of NIR spectroscopy for AD monitoring had already been evaluated several times, as shown for instance in Madsen et al. (2011). The starting position for this study was therefore the following: (a) There is nothing novel as such in conducting quantitative NIR measurements in the AD process; (b) Constituent intercorrelation is likely to compromise the integrity of the results; (c) Performing numerous digestions in a laboratory environment is very time-consuming; (d) Limited resources are available for carrying out reference measurements; (e) The concentration of the constituents of interest is very low, whereas absorption of IR light caused by water is very high. A potential solution to these issues was therefore found in the spiking of process samples. As mentioned in the Introduction, spiking has been used before in the context of bioprocess monitoring using IR spectroscopy. However, the present study is probably the first in which multi-constituent spiking is used, and, in addition to this, according to an (optimal) central composite design. This type of design is described (for instance) in Erikson et al. (2008).

In this thesis, two objectives were set for the spiking procedure. The first objective was to reduce constituent intercorrelation in the data. The second was to increase calibration model accuracy. These two objectives are in fact in contradiction to each other. By removing correlation that is useful from a mathematical point of view, the task of creating a model with seemingly higher accuracy becomes more difficult, in the sense that any correlation between whatever information is present in the spectra and the constituent of interest can be used in constructing a model. Thus, breaking intercorrelations can result in a model with seemingly lower accuracy. In order to address this aspect, the models regressed on process samples and the models regressed on spiked samples were also validated against the pure component spectra collected. Regarding the objective of reducing the constituent intercorrelation, the objective was fully met. For the constituents in the design, the intercorrelation was reduced (depending on the constituent in question) from around 0.5 to at least 0.02 in absolute values. In the case of model validation on pure component spectra, the coefficient of determination was between 2.4 and 12 times higher for the models regressed on spiked data. Thus, the models based on spiked data clearly contained features much more closely related to the actual spectral features of the constituent in question.

Evaluating whether the second objective was met, i.e., improving model accuracy, is far more difficult to prove. This is due to (among other things) the almost unlimited combinations of parameter settings and variable selection alternatives available in most PLS calibration software. Here, it was decided to try to find the best calibration model for each data set, rather than to specify a given combination of settings and then compare the model results. This somewhat compromises the evaluation of whether the second objective was met, but it was seen as the best alternative regardless. In any case, the results obtained state that the measurement error was reduced by at least 20% for all constituents studied when comparing models regressed on spiked samples to models regressed on process samples only.

In summary and in retrospect, this study did not represent a perfect example of the usefulness of orthogonal spiking for calibration purposes in bioprocess monitoring using IR spectroscopy. The concentration ranges of the constituents of interest were on the very limit of what can be measured with the equipment used. This was clearly indicated by the results from calibration models based on pure component spectra. Furthermore, the spike size used was too large, and widened the constituent calibration interval to an extent that was detrimental for the model accuracy. This study was carried out by comparing a large number of performance indicators to each other. Since all these details are found in Publication III, they will not be repeated here. Nevertheless, from a qualitative perspective, the following can be concluded. Clearly, although not surprisingly, orthogonal spiking can be used to reduce constituent intercorrelation. Apparently, or at least in this case, using spiked samples for model regression purposes results in models that rely on features more closely related to the actual features of the constituent of interest compared to models based on process samples. This should imply that the model is more reliable. Using the spiked calibration data also increased the accuracy for each of the constituents. Furthermore, preparing the spiking solution and spiking the samples took two days; performing a single batch digestion took three weeks. Without a single additional reference measurement, 33 calibration points were converted into 462 calibration points. As such, this basic methodology has the potential to be implementable in numerous other applications. Finally, the accuracy obtained for each constituent should be of interest to operators of anaerobic digestion plants, since all the measured constituents have been identified as very important for the successful implementation of AD (see Chapter 1).

3.3 NIR and MIR in comparison for anaerobic digestion measurements

Although some examples can be found, e.g., Di Egidio et al. (2010) and Sivakesava et al. (2001), simultaneous use of NIR and MIR spectroscopy for bioprocess monitoring is not that common. In the context of bioprocess monitoring, NIR and MIR have been called rival technologies, and it has been concluded that "The intrinsic advantages of MIR including the much broader wavelength range, and the fact MIR absorbance's are based on fundamental vibration modes of molecules, not only project MIR as the most powerful and significant technique in bioprocess monitoring but also relatively better than its rival NIR." (Roychoudhury et al., 2006). Whether or not this is the case will not be discussed in this thesis. This is simply a comparison of the accuracy obtained on spiked AD samples using NIR and MIR instruments. However, it is still comprehensive in the sense that it compares the performance of two instruments on four constituents, with or without sample preparation, on spiked or process samples, while trying to block out the human factor.

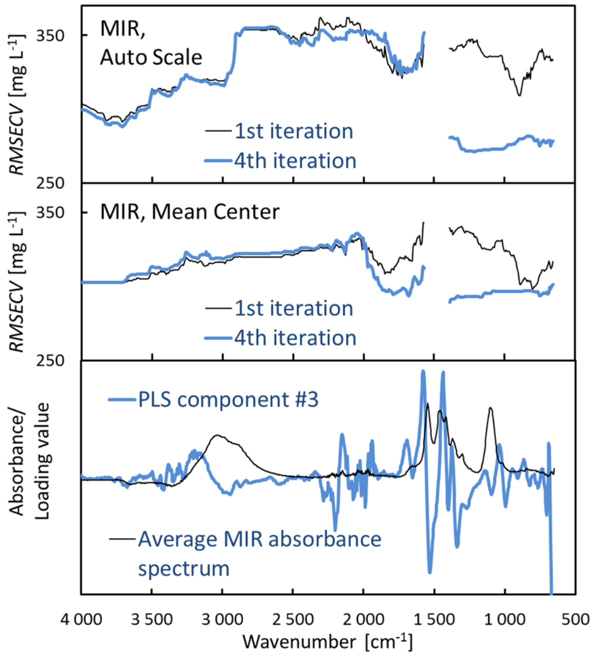


Figure 3.7. Impact of variable selection on *RMSECV* with MIR for acetate using auto scale and mean centre as pre-processing, sub-plotted with an average MIR spectrum and the third PLS component.

The impact of the variable selection procedure for the constituent acetate is illustrated in Figure 3.7 for MIR and Figure 3.8 for NIR. In the case of the MIR measurements in general, this procedure favoured the use of essentially all the collected spectral interval. However, this was not the case for the NIR measurements. As can be seen in Figure 3.8, the use of the higher wavelengths, particularly in combination with mean centring as pre-processing, resulted in a significantly higher *RMSECV*. This can be explained by the noisy signal in this region. The figure also shows that including the visible spectral information did not result in improved cross-validation predictions. The information in the PLS component was used to determine the initial interval to be used in the variable selection procedure, according to the apparent location of most of the relevant information in the PLS component.

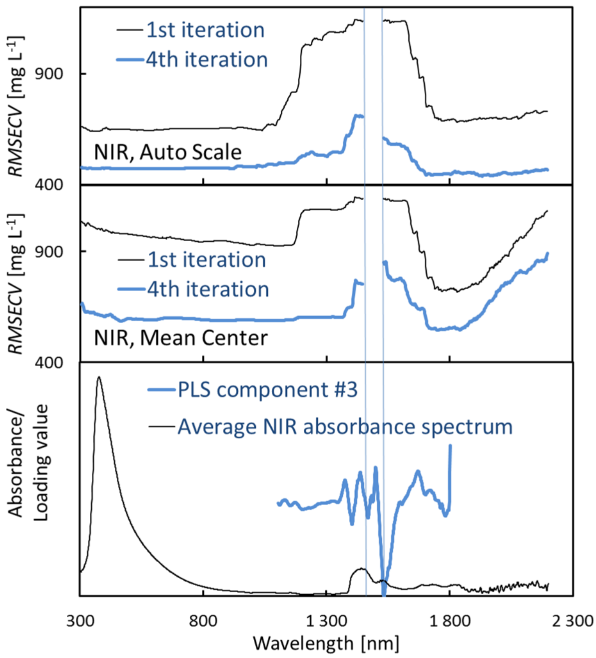


Figure 3.8. Impact of variable selection on *RMSECV* with NIR for acetate using auto scale and mean centre as pre-processing, sub-plotted with an average NIR spectrum and the third PLS component.

Figure 3.9 shows the impact of the spectral pre-processing on the *RMSECV* values. The figure reveals that the pre-processing methods used have only minor impacts on the *RMSECV* for the MIR calibrations. This seems reasonable. Creating calibrations for MIR spectroscopy should be a straightforward process. In contrast, switching from mean centre to auto scale in NIR has a dramatic impact. This is also the case for the number of points used in the polynomial when using a derivative as pre-processing with NIR in combination with auto scaling. On these specific data sets, NIR then produces a lower calibration error than MIR, provided that the proper pre-processing combinations are identified. Thus, these results suggest that creating an accurate quantitative calibration model for NIR is more demanding, both in terms of variable selection and in terms of selection pre-processing methods, than for MIR. Without presenting any further proof, it seems reasonable to assume that this is generally the case. This can perhaps be mainly attributed to the fact that peaks can usually be assigned to specific constituents in MIR, whereas the peaks in NIR are often broad and overlapping, originating from overtones and combination bands (Cooper et al., 1997).

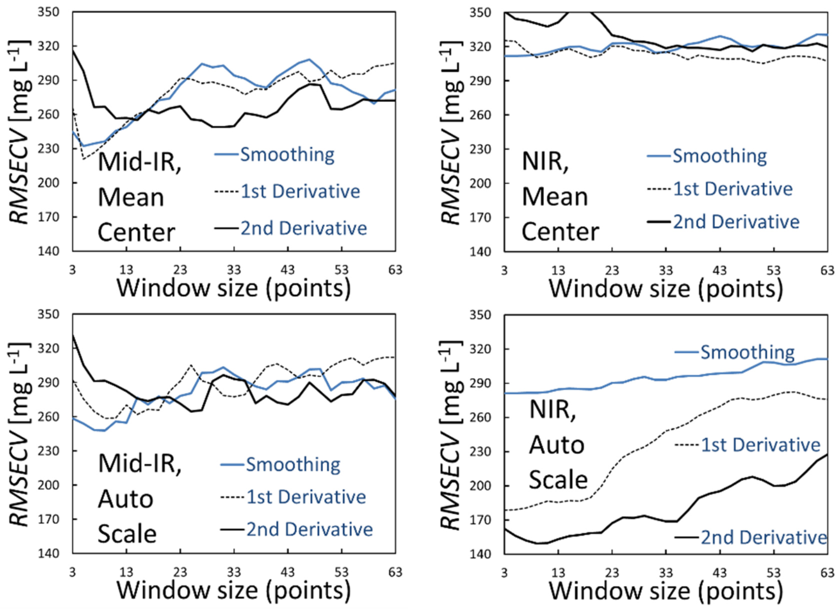


Figure 3.9. Impact on *RMSECV* (in propionate calibrations) of using smoothing, first order derivative, or second order derivative as spectral pre-processing, in combination with mean centre or auto scale for both MIR and NIR.

The *RMSEP* values for acetate and ammonium are displayed in Figure 3.10, and for propionate and TVFA in Figure 3.11. The arrows in these figures illustrate the impact of the measurement application alternatives studied: In Figure 3.10 (left side), in the form of how the instrument used impacted the measurement accuracy; in Figure 3.10 (right side), how the concentration range impacted the measurement accuracy; and in Figure 3.11 (right side), how the centrifugation of the samples impacted the measurement accuracy. One clear advantage with the ATR technique combined with MIR in bioprocess monitoring is that the presence of the solid phase should not affect the spectra. In contrast, in this study, a significant portion of the NIR spectra of unprepared samples had to be removed from the data sets due to the significant impact of the solid phase. The results depicted in Figure 3.10 and 3.11 seem to suggest that the sample preparation did not affect the NIR results that much, but these results should be viewed in light of the fact that 131 of the 924 spectra collected with NIR on unprepared samples had already been removed before these results were obtained. However, in the case of MIR, no spectra had to be removed due to the impact of the solid phase, and no benefits from centrifuging the samples are evident from the figures.

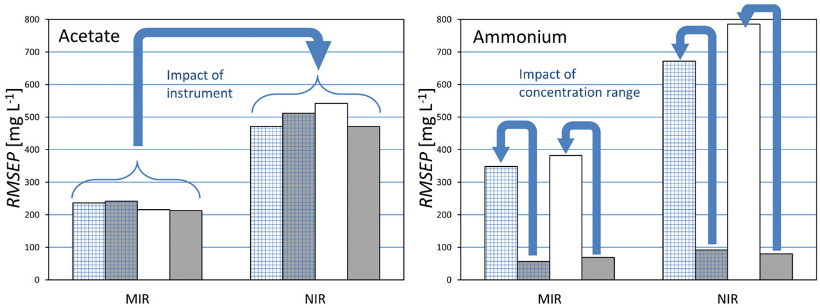


Figure 3.10. *RMSEP* values obtained for acetate and ammonium with MIR and NIR, for centrifuged (grid pattern) and unprepared (no pattern) samples, in the full concentration range (white fill) and the lower concentration range (grey fill).

In the case of ammonium, the concentration range had a dramatic impact on the *RMSEP* values for both MIR and NIR. The reason for this is not understood, but it clarifies why the use of local models, as described in Publication IV, had such significant effects on the measurement accuracy of ammonium. It is also interesting that the concentration range had a significant impact on the accuracy for propionate with NIR as well, whereas it had only a minor impact for propionate when using MIR. For TVFA, an impact of the sample preparation can be observed for NIR, but since so many spectra collected from unprepared samples were omitted before the comparison, no further conclusions can be drawn from this observation. Here, the NIR TVFA measurements on the centrifuged samples are actually more accurate than the corresponding ones with MIR. One explanation for this could be that since the TVFA concentration is the sum of many different VFA concentrations, the less specific information in the NIR spectra could actually be beneficial in comparison to MIR when constructing the calibration model. However, the *RMSEP* for the full concentration range on unprepared samples with MIR is nevertheless the most accurate model obtained for TVFA.

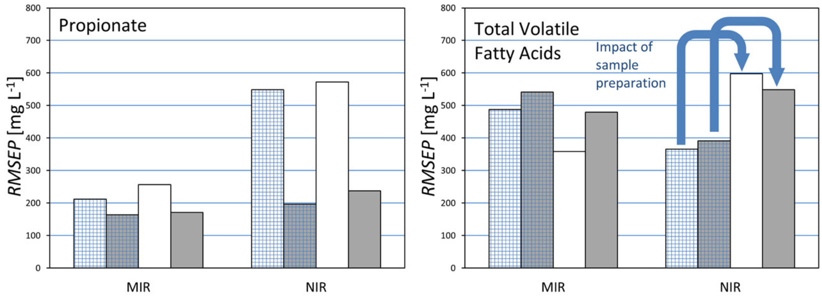


Figure 3.11. *RMSEP* values obtained for propionate and TVFA with MIR and NIR, for centrifuged (grid pattern) and unprepared (no pattern) samples, in the full concentration range (white fill) and the lower concentration range (grey fill).

It can be suggested that many of the observations made in this study also have broader relevance in terms of selecting between NIR and MIR spectroscopy for quantitative bioprocess monitoring. Unless information about the solid phase is of interest and can be obtained, the ATR technique is very convenient. In this actual comparison, ATR would in fact be practically the only option for true on-line analysis. However, if on-line measurements can be replaced by frequent off-line measurements on centrifuged samples, NIR could still be favourable from an economic point of view. As shown, obtaining a reliable model is more demanding with NIR than with MIR. This was illustrated by the impact of variable selection and pre-processing alternatives shown in this study, and it is likely that this is commonly the case. As such, it seems unlikely that NIR transmittance measurements would prove to be more accurate than MIR ATR measurements in liquid phase bioprocess monitoring. However, other considerations might still favour the use of NIR. Hence, NIR and MIR should be seen as complementary techniques, rather than rival ones, when it comes to bioprocess monitoring.

3.4 Local calibration methods and development of ML-PLS

The ML-PLS concept was first presented as a poster at the 14th International Conference of Near Infrared Spectroscopy (NIR 2009) in Bangkok, Thailand, and was later also published in the proceedings of that same conference by Dahlbacka & Lillhonga (2009). The application studied was moisture measurement in timber, and the incitement came from a pre-study on moisture measurement in biofuels using NIR spectroscopy, published by Dahlbacka (2010), in which it was suggested that the accuracy could be improved by using local models. The basic idea was that a global quantitative PLS model could be used to select which local quantitative model to use for the local model prediction. Here, the term local model meant that the model covered only a part of the constituent range studied. Further improvement in measurement accuracy was assumed to be obtainable by using the prediction of this local model to select a new local model regressed on an increasingly narrow moisture content range.

The model regression and validation methodology using this first version of ML-PLS is depicted in Figure 3.12 and 3.13. The approach and implementation at this point could be described as fairly primitive. The local models on each layer shared one moisture content point with its neighbour (on each side when applicable). The design presented consisted of four layers. All models had the same parameter settings, and all were used to predict the full validation data set. An Excel spreadsheet with if-statements was used to select which prediction to use on each layer. The prediction selection was based on the model's moisture content range into which the prediction on the previous layer fell. This four-layer ML-PLS model reduced the *RMSEP* from the 1.2% of the global model to 0.8%. However, at this point, the (weak) robustness of ML-PLS had already been identified as an issue, in the sense that the measurement accuracy was significantly impaired by a few highly faulty predictions. As a consequence, this improvement in measurement accuracy was obtained only after limiting the maximal difference in quantitative predictions between each layer.

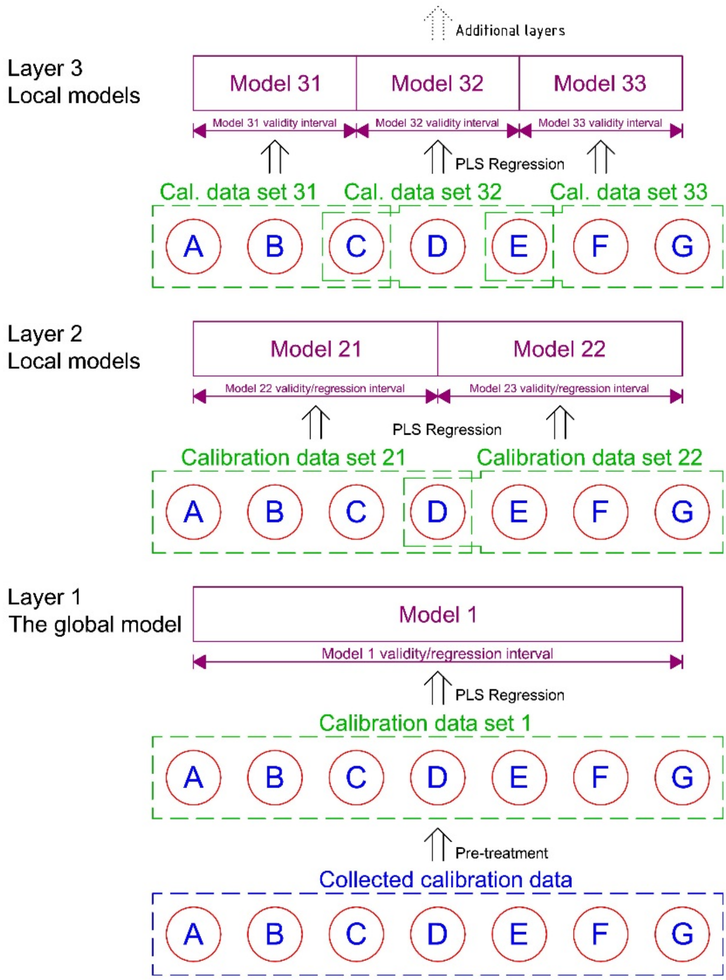


Figure 3.12. ML-PLS model regression scheme in the first version presented.

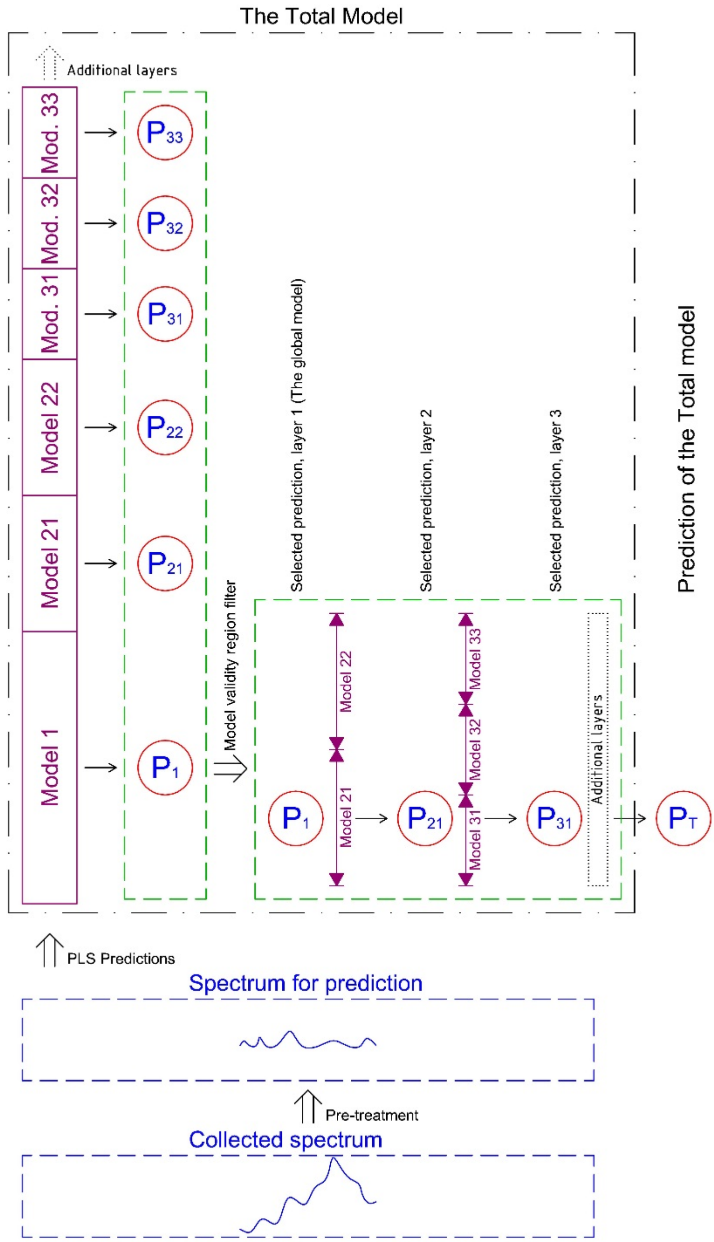


Figure 3.13. ML-PLS model prediction scheme in the first version presented.

The relative success of ML-PLS as described above, despite the primitive approach and implementation, encouraged further development of the method. However, as mentioned above, ML-PLS apparently had the tendency to make many highly accurate predictions of the validation data, but at the same time a few highly faulty predictions. It was argued that this was due to the following scenario: At some point in the design, a prediction error results in the selection of the wrong model on the next layer. The selection of the wrong model implies that the model is predicting a spectrum outside the range for which the model has been built. This will result in an even larger prediction error, which in turn will result in the selection of an even less-suited model on the next layer, and so forth. Thus, at higher layers in the design, predictions are sometimes made by models that are far out of their calibration range. In order to overcome this problem, a second phase in the development of ML-PLS was carried out by studying the same measurement application as before, i.e., moisture measurement in timber. Additional measurements were carried out and ML-PLS was implemented as a MATLAB script. The lack of robustness was addressed with the use of overlapping intervals, which in this case meant that local models on any given layer could share more than one moisture content point with other neighbouring models on the same layer.

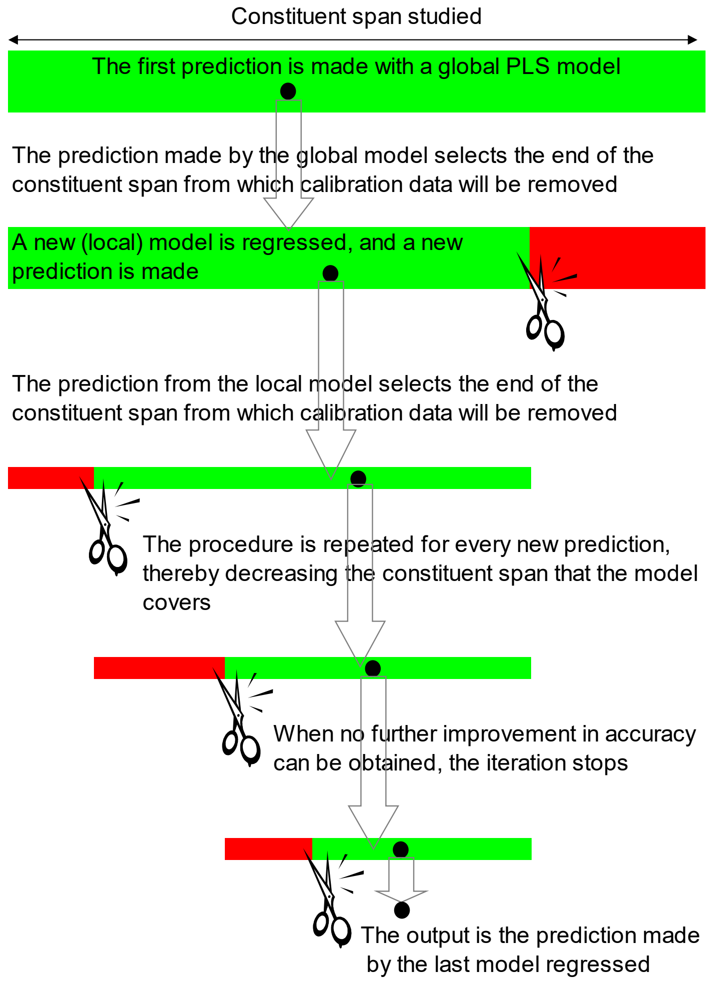


Figure 3.14. ML-PLS model prediction scheme as implemented in the first MATLAB script.

The conceptual way in which the model prediction in ML-PLS was implemented as a MATLAB script is shown in Figure 3.14. In practice, the script was specially adapted for the data set studied, and the design defined how many moisture content points should be removed from layer to layer. Overlapping intervals was achieved by simply removing less than 50% of the moisture content values from layer to layer. The use of overlapping intervals also introduced the need for a new model selection criterion. Whereas unique calibration intervals allowed for model selection based on into which model's interval the prediction from the previous layer fell, the selection was now based on which model's, or rather which local calibration data set's, median calibration interval was closest to the prediction on the previous layer. The impact of the design could be studied by removing a certain number of moisture content points in one step, making predictions on a design data set, computing the *RMSEP*, removing the same number of moisture content points in two steps, making predictions, computing the *RMSEP*, and comparing it to the previously-computed *RMSEP*. Using the methodology described above, an *RMSEP* of 1.16% was obtained with ML-PLS. In comparison, for the global model, an *RMSEP* of 2.11% was obtained using the same model parameter settings. However, with the global model, an *RMSEP* of 1.80% was also obtained for the best parameter settings found during the study. In any case, as shown in Dahlbacka & Lillhonga (2010), it was clear that ML-PLS could outperform a global PLS model on this data set. What was yet to be established was whether ML-PLS could match established local or non-linear calibration methods in terms of prediction accuracy.

In order to evaluate this, a comparison in the prediction accuracy on the same data set used in Dahlbacka & Lillhonga (2010) was made between LWR, support vector machine regression (SVMR, see Chen et al., 2005) and ML-PLS by Dahlbacka & Lillhonga (2012). Using PLS model regression in LWR, PLS compression in SVMR, seven PLS components for all methods and a second order Savitzky-Golay derivative with a second order seven-point polynomial and mean centring pre-processing, *RMSEPs* of 1.17%, 1.39%, 1.71%, and 1.81% was obtained for ML-PLS, LWR, SVMR and the global PLS model respectively. These parameter settings had also been identified as close to optimal on the data set studied. Thus, it was clear that local calibration techniques could significantly improve the measurement accuracy on this data set. It was also indicated that the ML-PLS method could perform as well as LWR. However, no significant improvement in measurement accuracy was found in this case by using non-linear methods in the form of SVMR. Figure 3.15 shows how the number of spectra in the regression data set or split affected the *RMSEP* value.

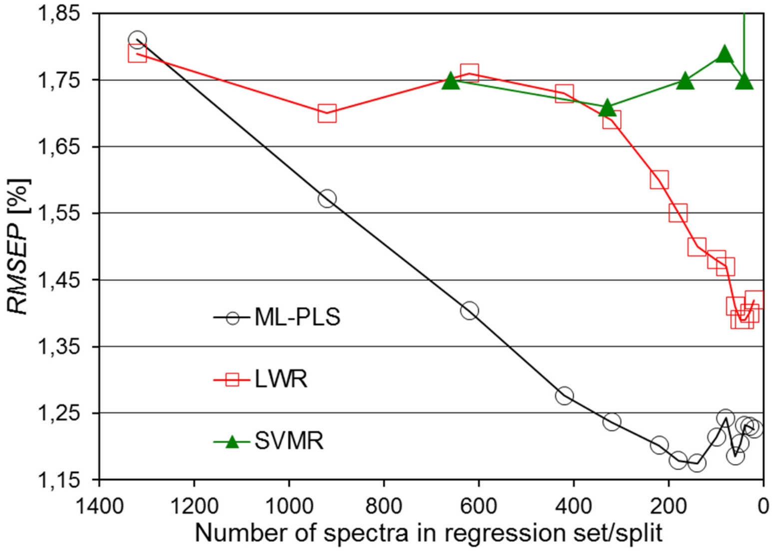


Figure 3.15. Impact of the number of spectra in the regression data set or split on the *RMSEP* value for ML-PLS, LWR and SVMR.

Up to this point, local calibration data subset or model selection in ML-PLS had been based only on the quantitative prediction of the model on the previous layer. However, as such, the iterative approach of ML-PLS could be utilised regardless of the type of selection method or distance measure selected. This being said, there are of course also static distance measures, e.g., spectral correlation, which on their own render the concept of an iterative approach pointless. In any case, testing other distance measures seemed to be a reasonable next step in the development of ML-PLS. Publication IV in this thesis evaluates the usability of NIR spectroscopy for quantitative measurements of ammonium and acetate in an AD process. It addresses three questions: (1) Are local calibration methods useful when implementing quantitative measurements in an AD process using NIR spectroscopy in the form of LWR and ML-PLS and in comparison to a global PLS model? (2) Are distance measures other than the distance to prediction used in the previous version of ML-PLS more effective in this application? (3) How does ML-PLS compare to the established local calibration method LWR?

This publication was made possible by the rapid means of generating calibration samples presented in Publication III. The study also illustrates the difficulties when comparing calibration methods that arise from the almost unlimited number of possible combinations of parameter settings when performing multivariate model regression. The distance measures studied in Publication IV were the Euclidean distance to prediction, Euclidean distance in spectral space, Euclidian distance in PLS space, absolute distance in spectral space, absolute distance in PLS space, Mahalanobis distance in PLS space, correlation in spectral space, correlation in PLS space, and the regression residual. The last one of these should perhaps not be described as a distance measure, because it simply removes the calibration data with the highest residuals regardless of the spectrum predicted. However, in the sense that it iteratively removes data from the local calibration data set, it is listed together with the other distance measures in the publication. A Euclidean distance d_1 between a prediction sample and a calibration sample is computed according to

$$\boldsymbol{d_1^2 = \sum_{a=1}^{A}(z_a - z_{pa})^2} \tag{3.1}$$

where z_a may be the spectral measurements of a calibration sample, the principal component scores of a calibration sample, or the PLS component scores of a calibration sample. Similarly, z_{pa} denotes the corresponding parameters of the prediction sample. A Mahalanobis distance d_2 between a prediction sample and a calibration sample is computed according to

$$\boldsymbol{d_2^2 = \sum_{a=1}^{A}\frac{(\hat{z}_a - \hat{z}_{pa})^2}{\hat{\lambda}_a}} \tag{3.2}$$

where $\hat{z}_a$ may be the principal component scores of a calibration sample, or the PLS component scores of a calibration sample, and $\hat{\lambda}_a$ is the eigenvalue of the principal or PLS component a. Similarly, $\hat{z}_{pa}$ denotes the corresponding parameters of the prediction sample.

In Publication IV, the risk of drawing the wrong conclusions when comparing calibration methods, due to the vast combination of possibilities of model parameters, was reduced by making numerous calibrations. In the case of LWR, 150 models were regressed using different parameter settings. These settings were quantitative predictions based on PLS, PCR and Global PCR, the size of the local subset of calibration spectra, the relative weight of the prediction in local subset selection, and the pre-processing method used. In practice, this was done by writing a MATLAB script for this specific purpose. The best result obtained for ammonium using LWR on a reduced validation data set was an *RMSEP* of 489 mg L^{-1}. In comparison, using the same spectral pre-processing, the global model generated an *RMSEP* of 897 mg L^{-1}. Thus, it is safe to say that LWR performed excellently on this data set for this constituent. However, Table 1 in Publication IV also reveals that the accuracy is highly dependent on the model parameter settings, further highlighting the dilemma of comparing multivariate calibration methods. By simply using manual (and thereby often random) testing of the impact of LWR compared to a global PLS model, this specific combination of model parameters that generated this low *RMSEP* for LWR would probably not have been found.

For ML-PLS and ammonium, the best distance measure found (on a reduced validation data set) was the distance to prediction. This was on mean centred data, and the *RMSEP* obtained was 186 mg L^{-1}, corresponding to 22% of that of the global model. However, on auto-scaled spectra, the Euclidian distance in PLS space was the most effective distance measure, resulting in an *RMSEP* of 296 mg L^{-1}, corresponding to 36% of that of the global model. An attempt was also made to use two distances (the distance to prediction and the Mahalanobis distance in PLS space) simultaneously, but at least on this small test, no improvement in accuracy was obtained this way.

The impact of the iterative procedure was investigated as well. It was concluded that on these data, the number of iterations or the step size had an impact on the measurement accuracy. Thus, using an iterative approach for the reduction of the calibration data set seemed favourable. The design was also "optimised" by studying the impact of the step size on *RMSEP* at different layers. It was shown that optimisation of the design had an impact on the measurement accuracy, and that this can also be useful in terms of minimising the number of layers in a design and thereby increasing the speed of prediction. The predictions of ammonium made by the global PLS model, the best LWR model found and layer 30 in the optimised ML-PLS model, are shown in Figure 3.16.

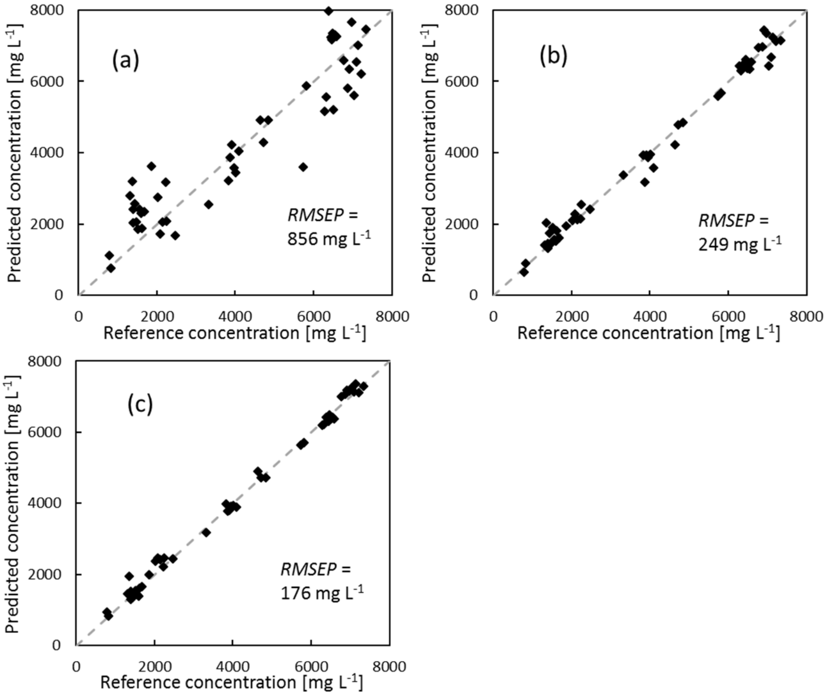


Figure 3.16. The predicted ammonium concentration plotted against the reference concentration on the reduced ammonium validation data set for (a) the global PLS model, (b) the best LWR model found and (c) layer 30 in an optimised ML-PLS model (45° line dashed).

As was the case when determining moisture content in timber as previously described, the ML-PLS prediction accuracy in Publication IV was again impaired on the full validation data set by a few apparent outliers with very high prediction residuals. This was particularly obvious when determining the ammonium concentration using an ML-PLS model with the Euclidian distance to prediction as the distance measure, already in the first design which was tested on the full validation data set. In agreement with Grubbs' (1969) definition, these poor predictions were classified as outliers, or rather as apparent outliers, because the reason for the poor predictions was not understood. Here, a significant amount of time was spent on trying to identify why these particular samples become outliers in ML-PLS. Among the methods tested to find these apparent outliers in the global PLS space and with the global PLS models were Hotelling's T^2 and Q residual (see Wise et al. (2006)), as well as prediction residuals or scores on the PLS components. However, on the level of the global data set and the global PLS model, these apparent outliers could not be detected as actual outliers. Other methods for outlier detection were also evaluated, but the only method that showed any promise was replicate analysis.

While the actual reason for these high residual predictions could not be identified, it was shown that the emergence of most of the apparent outliers was dependent on the distance measure used. Therefore, a very high accuracy could be obtained by comparing the predictions of two different ML-PLS models, one robust model and one mostly very accurate model. By stating that the prediction of the accurate model could only diverge a certain amount from the prediction of the robust model, and using the prediction of the robust model when the threshold was exceeded, an *RMSEP* of 274 mg L^{-1} was obtained with ML-PLS. This can be compared to the *RMSEP* of 897 mg L^{-1} with the global model. This was truly a remarkable improvement in measurement accuracy, although the way it was obtained was certainly not very straightforward. Figure 3.17 shows the ammonium prediction results on the full validation data set with the models described here. As a sharp contrast to the improvements in the prediction accuracy obtained for ammonium using local models, essentially no improvement in the measurement accuracy of the acetate concentration was obtained with either of the local methods. This being said, the improvement in accuracy obtained for ammonium in Publication IV still motivated further development of the method.

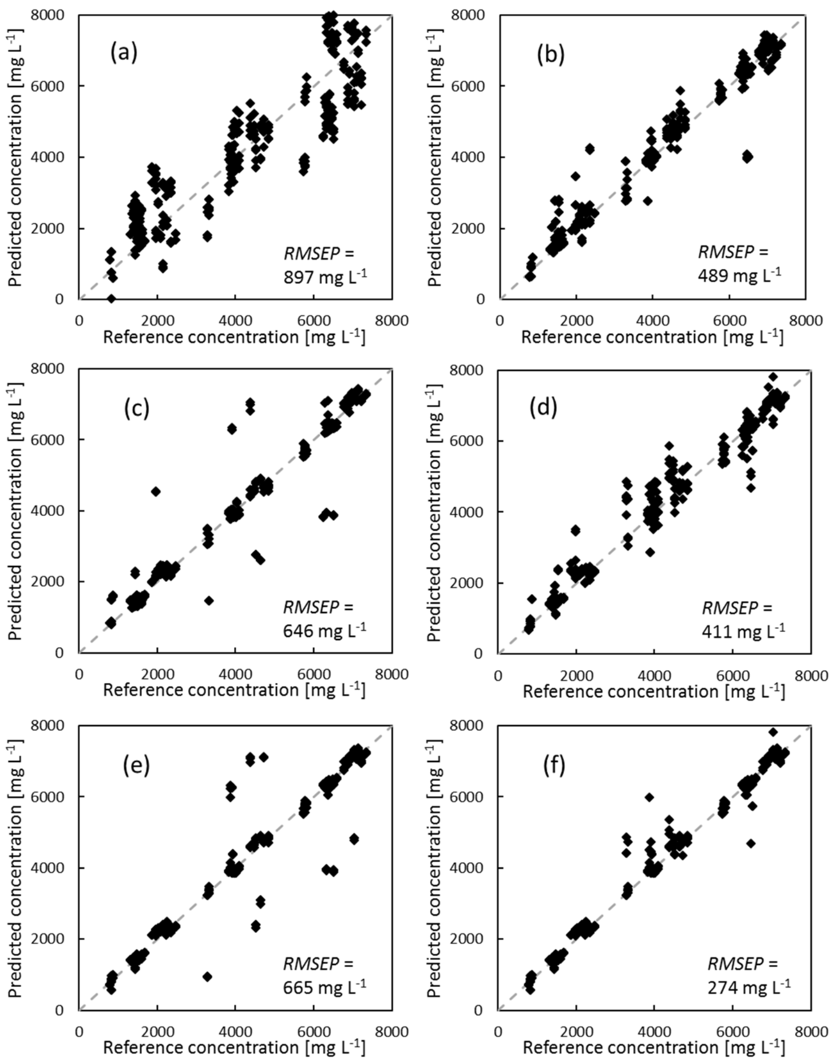


Figure 3.17. The predicted concentration plotted against the reference concentration on the full ammonium validation data set for (a) the global PLS model, (b) the best LWR model found, layer 30 in the optimised ML-PLS design using (c) the Euclidean distance to prediction as selection method and mean centring as spectral pre-processing, (d) the Euclidean distance in PLS space as selection method and auto scale as spectral pre-processing, (e) a combination of these two selection methods and (f) a threshold-based combination of the predictions shown in (d) and (e) (45° line dashed).

The latest version of ML-PLS is explained in some detail in Publication V of this thesis. However, for clarity, it will also be described here together with some supplementary information. At this point of the development of ML-PLS, the reasoning was as follows: The ML-PLS concept seems to work. There is a need for automated creation of (optimal) designs. Additional distance measures might as well be included, and in order to promote the use of ML-PLS, a graphical user interface (GUI) should be created. Thus, this is what was done. It seems suitable to use the screen shot of the ML-PLS GUI shown in Figure 3.18 as the basis for the methodology description. For further information, the actual flowchart of ML-PLS regression is shown in Figure 1 of Publication V. Starting from the top-left part of the screen shot, calibration data can easily be loaded from the workspace. This is also the case for separate sets of design and validation data. However, the design data may also be extracted or copied automatically from the calibration data. When the data is extracted, it is removed from the original calibration data set and thereby becomes a fully external data set. If the design data is instead copied, the same data will also still be used for calibration purposes. This latter option can be useful when there is a limited amount of calibration data available. However, building the design on the same data as the calibration data also comes with a risk similar to that of overfitting.

Figure 3.18. Screen shot of the GUI for ML-PLS.

The last design created during the current session can be validated by a "push of a button". Alternatively, previously-saved designs can be reloaded and validated. The pre-processing options and selection interface are the same as in the PLS Toolbox, and the same algorithms for PLS and PCA are available. The design can be built by selecting the distance measure that gives the lowest *RMSEP* when moving through the layers, or by selecting the distance measure that gives the lowest median prediction error. The latter option is considered a robust alternative, essentially intended to remove the effect of outliers in the design data set. The extraction or copying of design data from the calibration data set can be carried out either randomly or as equidistant blocks, also known as Venetian blinds. The amount of data to be used as design data is specified as a percentage of the full calibration data set.

The creation of a design is terminated in one of three possible ways: (1) when the specified maximal number of layers has been reached, (2) when the optimisation criterion (*RMSEP* or median prediction error) has increased for a specified number of layers, or (3) when the maximal number of PLS or PCA components exceeds the number of remaining local calibration spectra. Step size optimisation is carried out for the specified number of step size divisions. If the number of step size divisions is set to zero, no step size optimisation takes place. This is also always the case for the static distances, i.e., the distances in spectral space. This version of ML-PLS supports the use of 12 distances or distance measures, listed in Table 3.1. The table also includes the absolute Y-residual (*yres*), i.e., the largest cross-validation residuals on the calibration data, although this is not technically a distance to the unknown spectrum.

Table 3.1. Distance measures and abbreviations of these used in ML-PLS and the ML-PLS GUI respectively.

Distance measure	Abbreviation
Euclidean distance to prediction	*epred*
Absolute distance in spectral space	*adiff*
Euclidean distance in spectral space	*ediff*
Correlation in spectral space	*cdiff*
Absolute distance in PLS score space	*adpls*
Euclidean distance in PLS score space	*edpls*
Correlation in PLS score space	*cdpls*
Mahalanobis distance in PLS score space	*mdpls*
Absolute distance in PC score space	*adpca*
Euclidean distance in PC score space	*edpca*
Mahalanobis distance in PC score space	*mdpca*
Absolute Y-residual	*yres*

The creation and optimisation of a design takes place as follows. This description is valid when the *RMSEP* value is the target value. The target value can also be the median absolute prediction deviation, but in order to simplify this description, this option is omitted in this explanation. First a global PLS model is regressed. The number of PLS components to be used in the model (and all local models as well) is selected according to the number of PLS components that gives the lowest *RMSECV* value. The design data set is thereafter predicted. For a given spectrum in the design data set and for a given distance measure, a local calibration data set is extracted from the calibration data set. Here the Euclidean and Mahalanobis distances are computed as truncated values, according to the number of PLS components in use on the previous layer or in the previous iteration. A PLS model is thereafter regressed and a prediction is made on the design spectrum.

This procedure is repeated for all distance measures in use and thereafter for each spectrum in the design data set.

When this has been carried out, an *RMSEP* is computed for each distance measure based on the predictions of the whole design data set for this given distance. The distance corresponding to the lowest *RMSEP* is thereafter selected as the distance to be used in the transition from the previous layer to this layer. If the number of step size divisions is greater than zero and unless the distance measure to be used is a static distance measure, step size optimisation will now take place. For the selected distance, the step size is cut in half. The removal of the same amount of data as with the previous step size is thereafter performed in two steps. A new *RMSEP* is computed. If it is lower than for the previously-used step size, the halved step size is now selected. If this happens, the procedure is repeated for the specified number of step size divisions or until the reduction of the local data set in two steps instead of one results in an increase in the *RMSEP*.

In all, the described procedure will result in an "optimal" design. The design is essentially a description of which distance measure and step size to use on each layer. The number of layers to be used for predicting an unknown spectrum or to validate the model is given as the layer with the lowest *RMSEP* on the design data. As the description of the construction of a design might suggest, this can be very calculation-intense. However, if there is a need for faster design construction, this can readily be done, for instance, by reducing the size of the design data set, excluding distance measures, or increasing the step size. Making a prediction using a standard laptop should in any case take less than a minute, and this prediction speed is certainly fast enough for more or less any bioprocess application.

The methodology and script presented above was evaluated in Publication V on four very different data sets. However, only one of these data sets originated from bioprocess measurements. Thus, the results from Publication V are described only briefly. Besides the new methodology used, the main point of Publication V was to evaluate the usefulness of ML-PLS as such. This was done by comparing the accuracy obtained with ML-PLS to that of a global model and accuracies reported in the literature on the same data sets. The NIR spectroscopy data and constituent values set used in this study were collected from soil samples (Brown et al. 2006), single wheat kernels (Nielsen et al., 2003), diesel fuel (collected by Southwest Research Institute, no reference available), and spiked AD samples (Publication III).

The first data set studied came from soil samples. In this case, the clay content was selected as the constituent of interest and the constituent to be studied (from a number of potential constituents available in the data set). This soil sample data set was in many ways perfect for studying the usefulness of ML-PLS. It was large (4184 spectra), and in addition to standard PLS regression, other calibration methods had already been applied to it in the form of boosted regression trees (BRT, see Brown et al., 2006), as well as SVM, LWR, LOCAL, and a spectrum-based learner (SBL) (Ramirez-Lopez et al., 2013). Allowing the use of all 12 available distance measures generated an ML-PLS model with an *RMSEP* of 88 g kg^{-1}. This value can be compared to the *RMSEP* of 111 g kg^{-1} of the global PLS model, the *RMSECV* of 95 g kg^{-1} with BRT (Brown et al., 2006), and the *RMSEP* of 120 g kg^{-1} (12.0%) using SBL (Ramirez-Lopez et al., 2013). Thus, in this particular case, ML-PLS performed very well according to all available means of comparison. Furthermore, this result was obtained with no supervision or interference in the ML-PLS model regression step – which was not the case for the other data sets.

The second data set studied was retrieved from single wheat kernels. This data set had also been studied in several publications, and in Publication V the protein content was chosen as the constituent of interest. When using ML-PLS, the *RMSEP* decreased to 0.38% percentage protein content in dry matter compared to an *RMSEP* of 0.48% for the global model and an *RMSEP* ranging between 0.42% and 0.48% using global PLS, kernel PLS (KPLS), support vector machines (SVM), least-squares SVM (LS-SVM), relevance vector machines (RVM), Gaussian process regression (GPR), artificial neural network (ANN), and Bayesian ANN (BANN) (as reported by Ni et al. (2013)). However, this accuracy for ML-PLS was obtained only after selecting the best pair of distance measures found.

In the case of the freezing point in diesel, a 30% reduction in the *RMSEP* was obtained in comparison to the global PLS model.

However, the size and characteristics of this diesel data set were less than ideal (see Publication V) for testing local calibration methods, so the result is thereby perhaps not that interesting.

In the case of the spiked AD sample data set, ammonium was selected as the constituent of interest since promising results had been obtained previously with ML-PLS for this constituent. As described in Publication IV, the potential increase in prediction accuracy was here again impaired by a few outliers. Some unsuccessful attempts were made to understand the reason behind these few poor predictions. Instead, increased robustness was obtained by using only three latent variables, *epred* and distances in PC and PLS space as distance measures, and a step size of 70 percent. With these settings, a reduction of *RMSEP* by 50% (corresponding to 407 g L^{-1}) was obtained with ML-PLS compared to the global PLS model. Figure 3.19 shows the predictions made of ammonium using either seven or three latent variables when using the step size and distance measures described above. Overall, this first study on the latest ML-PLS version indicated that ML-PLS can be a very powerful calibration method. However, it also indicated that the best distance measure on each layer does not necessarily result in the best possible ML-PLS model.

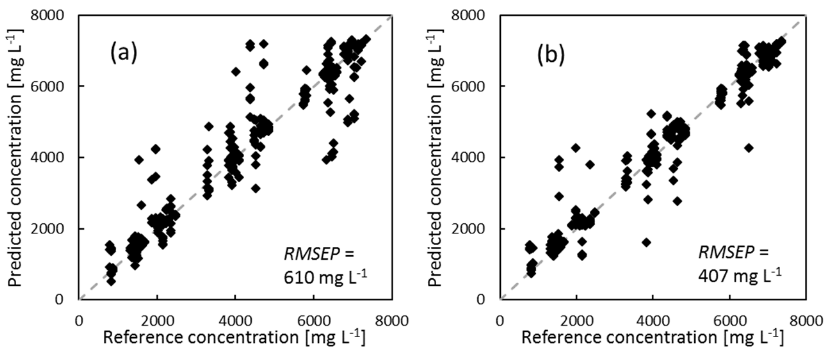


Figure 3.19. Predicted ammonium concentration on the last layer plotted against the reference concentration for the AD validation data set using a full step size of 70% and (a) seven latent variables, (b) three latent variables.

Although the ML-PLS concepts cannot yet be said to have undergone a thorough evaluation, it seems appropriate to point out some general observations about the method. One recurring observation is that a few outliers often make a significant impact on the measurement accuracy. This issue has not been fully solved yet; however, a few outliers are always statistically to be expected. Another potential issue, which has not been discussed here, is that the creation of designs and obtaining predictions is time-consuming, at least in comparison to traditional PLS model regression and prediction. However, whether or not this is an issue depends in the end on how quickly a new design must be created and how quickly a prediction must be made. In most bioprocess applications, obtaining a prediction roughly every minute using an ordinary laptop should be a sufficient sampling speed. Therefore, this methodology can hopefully make an impact on future applications of bioprocess monitoring using IR spectroscopy when there is an abundance of calibration data available, the spectral features are complex, and the constituent of interest is difficult to model with a global multivariate method.

Chapter 4

4. Summary and conclusions

This thesis describes the implementation of quantitative measurements of highly relevant constituents in bioprocesses. Measurement applications are implemented for *Pichia pastoris* and *Streptomyces peucetius* fermentations, as well as anaerobic digestion using MIR spectroscopy. It is clear that the measurements in the *Pichia pastoris* fermentation and in the AD process could potentially be very important when used for process control. Since the fermentations of *Streptomyces peucetius* were carried out as batch fermentations, the usefulness of the measurement application implemented is perhaps limited to increased process knowledge. However, this is of value in itself. A measurement application is also implemented for the AD process using NIR spectroscopy. Although examples of on-line measurements with NIR in the AD process by means of transflection spectroscopy can be found, it seems reasonable to suggest that on-line transmittance spectroscopy with NIR can be implemented only in combination with an automated filtration unit. Nevertheless, since this is a relatively slow process, off-line measurements may also be used for process control. Used in this way, NIR spectroscopy represents a fairly inexpensive measurement alternative.

Calibration aspects are addressed in the form of creation of calibration data using spectral addition, spiking of process samples, and by employing local calibration methods. The complex constituent matrix and low concentration range typically found in bioprocesses make it difficult to perform reliable quantitative IR spectroscopy. If the constituent of interest creates most of the variance in the spectral data in a situation where the signal to noise ratio is high, creating a quantitative calibration model should be trivial. In the best-case scenario, a single calibration spectrum and a single wavelength or wavenumber could be all that is needed to perform quantitative measurements. However, this is rarely the case in bioprocess applications. A sufficiently large and intercorrelation-limited data set in combination with multivariate methods is usually needed to implement a successful quantitative measurement.

Creating spectra by spectral addition is trivial. However, this is also the beauty of the method, representing a novel concept for bioprocess applications. Any single spectrum can be modelled perfectly with a single loading vector in PLS. However, it is the combination of spectral and constituent data that gives unlimited possibilities to manipulate the calibration data set. Simply including the same spectra twice in the calibration data will, for instance, give more weight to the information in and associated with this spectrum in the regression of the calibration model. In summary, perhaps the main contributions of this part of the thesis to the field of quantitative measurements in bioprocesses using IR spectroscopy is simply to increase awareness about the possibilities of using mathematical manipulations of the calibration data set.

The orthogonal multi-constituent spiking methodology presented can be described as a novel methodology that should be widely applicable in bioprocesses. In this thesis, the impact on the measurement accuracy was significant, particularly when taking into account that the spiking procedure reduces intercorrelation in the data and thereby makes calibration more difficult. In contrast to making simple mathematical manipulations of the calibration data set, this methodology actually results in a real and unique spectrum containing real chemical information for each spiked calibration sample. The main drawback with the method is probably the potential risk of creating chemical reactions in the samples due to the introduction of the spikes. To some extent, spiking also creates extra work compared to only using process samples, if the comparison is made for a given number of process samples with or without performing spiking. If the comparison is made between a given number of spiked samples and the same number of process samples, the situation should commonly be the reversed. Spiking reduces the time spent on reference measurements and production/collection of process samples. As a bonus, the spiked samples should basically always be characterised by a lower constituent intercorrelation compared to process samples (provided that the spiking has been done properly).

In the measurements on AD samples, promising results were obtained using local calibration methods in the form of LWR and ML-PLS. However, using local models for quantitative measurements in bioprocesses is certainly not a very common procedure. Moreover, it seems reasonable to assume that local calibration models will never be of any greater significance in the case of MIR spectroscopy on bioprocesses. However, this is not necessarily the case when it comes to NIR spectroscopy on bioprocesses. Both ML-PLS and LWR resulted in a dramatic increase of measurement accuracy for ammonium, but this was not the case for acetate. Creating local calibrations on spiked data also creates a certain danger of obtaining accuracy improvement by lumping together or extracting separate spikes into the local calibration data. In any case, it would be very interesting to see further evaluations on the usability of local calibration techniques in the context of bioprocess monitoring.

As many experts have stated, the importance of biotechnology is continuously growing, and bioprocesses are an important part of this. Thus, the ability to optimise bioprocess production through new measurement applications, by increasing process knowledge and enabling process control, will certainly be of great importance for many years to come. This being said, closed loop control of bioprocesses relying on quantitative measurements using IR spectroscopy is still very uncommon. However, in order to make full use of IR spectroscopy's potential, closed loop control is certainly a key objective. This, in turn, can only be achieved when reliable quantitative models are available. This thesis also addresses the creation of calibration models from different perspectives. Thus, yet another step is taken towards increasing the occurrence of closed loop control based on IR measurements in bioprocess operations, although this goal has yet to be reached.

References

Aastveit, A. H., & Marum, P. (1993). Near-infrared reflectance spectroscopy: different strategies for local calibrations in analyses of forage quality. *Applied spectroscopy*, *47*(4), 463-469.

Abu-Absi, N. R., Martel, R. P., Lanza, A. M., Clements, S. J., Borys, M. C., & Li, Z. J. (2014). Application of spectroscopic methods for monitoring of bioprocesses and the implications for the manufacture of biologics. *Pharmaceutical Bioprocessing*, *2*(3), 267-84.

Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. *Trends in Analytical Chemistry*, *21*(4), 240-250.

Beutel, S., & Henkel, S. (2011). In situ sensor techniques in modern bioprocess monitoring. *Applied microbiology and biotechnology*, *91*(6), 1493.

Biechele, P., Busse, C., Solle, D., Scheper, T., & Reardon, K. (2015). Sensor systems for bioprocess monitoring. *Engineering in Life Sciences*, *15*(5), 469-488.

Bluma, A., Höpfner, T., Lindner, P., Rehbock, C., Beutel, S., Riechers, D., Hitzmann, B. and Scheper, T. (2010). In-situ imaging sensors for bioprocess monitoring: state of the art. *Analytical and bioanalytical chemistry*, *398*(6), 2429-2438.

Bochenkov, V. E., & Sergeev, G. B. (2010). Sensitivity, selectivity, and stability of gas-sensitive metal-oxide nanostructures. *Metal oxide nanostructures and their applications*, *3*, 31-52.

Boe K. (2006). *On-line monitoring and control of the biogas process*. Institute of Environment and Resources, DTU. Ph.D. Thesis.

Brink, A., & Westerlund, T. (1995). The joint problem of model structure determination and parameter estimation in quantitative IR spectroscopy. *Chemometrics and intelligent laboratory systems*, *29(1)*, 29-36.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3), 273-290.

Cereghino, G. P. L., Cereghino, J. L., Ilgen, C., & Cregg, J. M. (2002). Production of recombinant proteins in fermenter cultures of the yeast Pichia pastoris. *Current opinion in biotechnology*, *13*(4), 329-332.

Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A., & Gernaey, K. V. (2009). Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnology progress*, *25*(6), 1561-1581.

Chen, P. H., Lin, C. J., & Schölkopf, B. (2005). A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, *21*(2), 111-136.

Clementschitsch, F., & Bayer, K. (2006). Improvement of bioprocess monitoring: development of novel concepts. *Microbial cell factories*, *5*(1), 19.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368), 829-836.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, *83*(403), 596-610.

Cooper, J. B., Wise, K. L., Welch, W. T., Sumner, M. B., Wilt, B. K., & Bledsoe, R. R. (1997). Comparison of near-IR, Raman, and mid-IR spectroscopies for the determination of BTEX in petroleum fuels. *Applied spectroscopy*, *51*(11), 1613-1620.

Crowley, J., McCarthy, B., Nunn, N. S., Harvey, L. M., & McNeil, B. (2000). Monitoring a recombinant Pichia pastoris fed batch process using Fourier transform mid-infrared spectroscopy (FT-MIRS). *Biotechnology letters*, *22*(24), 1907-1912.

Crowley, J., Arnold, S. A., Wood, N., Harvey, L. M., & McNeil, B. (2005). Monitoring a high cell density recombinant Pichia pastoris fed-batch bioprocess using transmission and reflectance near infrared spectroscopy. *Enzyme and microbial technology*, *36*(5), 621-628.

Cuetos, M. J., Gómez, X., Otero, M., & Morán, A. (2010). Anaerobic digestion of solid slaughterhouse waste: study of biological stabilization by Fourier Transform infrared spectroscopy and thermogravimetry combined with mass spectrometry. *Biodegradation*, *21*(4), 543-556.

Dahlbacka, J., & Lillhonga, T. (2009). Multi-Layer PLS Modeling as a Method to Master Model Stiffness – Applied to Moisture Measurement in Timber. Proceedings of NIR-2009, Breaking the Dawn. The 14th International Conference of Near Infrared Spectroscopy. Bangkok, Thailand, 7-16 November. 891-893.

Dahlbacka, J. (2010). Characterisation of Wooden Biofuels Using Near Infrared Spectroscopy – A Pre-Study. *Novia publications and Productions*, Series R: Reports, 4/2010.

Dahlbacka, J., & Lillhonga, T. (2010). Moisture measurement in timber utilising a multi-layer partial least squares calibration approach. *Journal of Near Infrared Spectroscopy*, *18*(6), 425-433.

Dahlbacka, J., & Lillhonga, T. (2012). A comparison of local and non-linear calibration methods for determination of moisture content in wood samples. Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa. 123-126.

Davies, A. M., Britcher, H. V., Franklin, J. G., Ring, S. M., Grant, A., & McClure, W. F. (1988). The application of fourier-transformed near-infrared spectra to quantitative analysis by comparison of similarity indices (CARNAC*). Microchimica Acta*, *94*(1), 61-64.

Dean, J. R. (2003). *Methods for environmental trace analysis* (Vol. 12). John Wiley and Sons.

Doyle, W. M. (1995). Near-IR and mid-IR process analysis-a critical comparison. *Advances in Instrumentation and Control, 50*(1), 433-441.

Di Egidio, V., Sinelli, N., Giovanelli, G., Moles, A., & Casiraghi, E. (2010). NIR and MIR spectroscopy as rapid methods to monitor red wine fermentation. E*uropean Food Research and Technology, 230*(6), 947-955.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C. & Wold, S. (2008). *Design of Experiments: Principles and Applications*. Umetrics Academy, Umeå.

Fearn, T. (2002). Assessing calibrations: SEP, RPD, RER and R2. *NIR news, 13*(6), 12-14.

Fearn, T., & Davies, A. M. (2003). Locally-biased regression. *Journal of Near Infrared Spectroscopy, 11*(6), 467-478.

Finn, B., Harvey, L. M., & McNeil, B. (2006). Near-infrared spectroscopic monitoring of biomass, glucose, ethanol and protein content in a high cell density baker's yeast fed-batch bioprocess. *Yeast, 23*(7), 507-517.

Franco, V. G., Perín, J. C., Mantovani, V. E., & Goicoechea, H. C. (2006). Monitoring substrate and products in a bioprocess with FTIR spectroscopy coupled to artificial neural networks enhanced with a genetic-algorithm-based method for wavelength selection. *Talanta, 68*(3), 1005-1012.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta, 185*, 1-17.

Gorry, P. A. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Analytical Chemistry, 62*(6), 570-573.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Guarna, M.M., Lesnicki, G.J., Tam, B.M., Robinson, J., Radziminski, C.Z., Hasenwinkle, D., Boraston, A., Jervis, E., MacGillivray, R.T.A., Turner, R.F.B. & Kilburn, D.G. (1997). On-line monitoring and control of methanol concentration in shake-flask cultures of Pichia pastoris. *Biotechnology and bioengineering*, *56*(3), 279-286.

Hansson, M., Nordberg, Å., Sundh, I., & Mathisen, B. (2002). Early warning of disturbances in a laboratory-scale MSW biogas process. *Water Science and Technology*, *45*(10), 255-260.

Hansson, M., Nordberg, Å., & Mathisen, B. (2003). On-line NIR monitoring during anaerobic treatment of municipal solid waste. *Water Science and Technology*, *48*(4), 9-13.

Hart, J. R., Norris, K. H., & Golumbic, C. (1962). Determination of the moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. *Cereal Chemistry*, *39*(2), 94-99.

Holm-Nielsen, J. B., Andree, H., Lindorfer, H., & Esbensen, K. H. (2007). Transflexive embedded near infrared monitoring for key process intermediates in anaerobic digestion/biogas production. *Journal of Near Infrared Spectroscopy*, *15*(2), 123-135.

Holm-Nielsen, J. B., Lomborg, C. J., Oleskowicz-Popiel, P., & Esbensen, K. H. (2008). On-line near infrared monitoring of glycerol-boosted anaerobic digestion processes: Evaluation of process analytical technologies. *Biotechnology and bioengineering*, *99*(2), 302-313.

Jacobi, H. F., Moschner, C. R., & Hartung, E. (2009). Use of near infrared spectroscopy in monitoring of volatile fatty acids in anaerobic digestion. *Water Science and Technology*, *60*(2), 339-346.

Jacobi, H. F., Moschner, C. R., & Hartung, E. (2011). Use of near infrared spectroscopy in online-monitoring of feeding substrate quality in anaerobic digestion. *Bioresource technology*, *102*(7), 4688-4696.

Jantsch, T. G., & Mattiasson, B. (2004). An automated spectrophotometric system for monitoring buffer capacity in anaerobic digestion processes. *Water Research*, *38*(17), 3645-3650.

Kjeldahl, K., & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, *24*(7-8), 558-564.

Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006, July). Detecting spam blogs: A machine learning approach. *In AAAI* (Vol. 6, pp. 1351-1356).

Krapf, L. C., Heuwinkel, H., Schmidhalter, U., & Gronauer, A. (2013). The potential for online monitoring of short-term process dynamics in anaerobic digestion using near-infrared spectroscopy. *Biomass and Bioenergy*, *48*, 224-230.

Landgrebe, D., Haake, C., Höpfner, T., Beutel, S., Hitzmann, B., Scheper, T., ... & Reardon, K. F. (2010). On-line infrared spectroscopy for bioprocess monitoring. *Applied microbiology and biotechnology*, *88*(1), 11-22.

Li, X., Dai, X., Takahashi, J., Li, N., Jin, J., Dai, L., & Dong, B. (2014). New insight into chemical changes of dissolved organic matter during anaerobic digestion of dewatered sewage sludge using EEM-PARAFAC and two-dimensional FTIR correlation spectroscopy. *Bioresource technology*, *159*, 412-420.

Lomborg, C. J., Holm-Nielsen, J. B., Oleskowicz-Popiel, P., & Esbensen, K. H. (2009). Near infrared and acoustic chemometrics monitoring of volatile fatty acids and dry matter during co-digestion of manure and maize silage. *Bioresource technology*, *100*(5), 1711-1719.

Lopes, J. A., Costa, P. F., Alves, T. P., & Menezes, J. C. (2004). Chemometrics in bioprocess engineering: process analytical technology (PAT) applications. *Chemometrics and Intelligent Laboratory Systems*, *74*(2), 269-275.

Lourenço, N. D., Lopes, J. A., Almeida, C. F., Sarraguça, M. C., & Pinheiro, H. M. (2012). Bioreactor monitoring with spectroscopy and chemometrics: a review. *Analytical and bioanalytical chemistry*, *404*(4), 1211-1237.

Madsen, M., Holm-Nielsen, J. B., & Esbensen, K. H. (2011). Monitoring of anaerobic digestion processes: A review perspective. R*enewable and Sustainable Energy Reviews*, *15*(6), 3141-3155.

Madsen, M., Ihunegbo, F. N., Holm-Nielsen, J. B., Halstensen, M., & Esbensen, K. H. (2012). On-line near infrared monitoring of ammonium and dry matter in bioslurry for robust biogas production: a full-scale feasibility study. *Journal of Near Infrared Spectroscopy*, *20*(6), 635-645.

Martínez, E. J., Fierro, J., Sánchez, M. E., & Gómez, X. (2012). Anaerobic co-digestion of FOG and sewage sludge: Study of the process by Fourier transform infrared spectroscopy. *International biodeterioration & biodegradation*, *75*, 1-6.

Milligan, M., Lewin-Koh, N., Coleman, D., Arroyo, A., & Saucedo, V. (2014). Semisynthetic model calibration for monitoring glucose in mammalian cell culture with in situ near infrared spectroscopy. *Biotechnology and bioengineering*, *111*(5), 896-903.

Nielsen, J. P., Pedersen, D. K., & Munck, L. (2003). Development of nondestructive screening methods for single kernel characterization of wheat. *Cereal chemistry*, *80*(3), 274-280.

Nordberg, Å., Hansson, M., Sundh, I., Nordkvist, E., Carlsson, H., & Mathisen, B. (2000). Monitoring of a biogas process using electronic gas sensors and near-infrared spectroscopy (NIR). *Water Science and Technology*, *41*(3), 1-8.T.

Naes, T., Isaksson, T., & Kowalski, B. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, *62*(7), 664-673.

Næs, T., & Isaksson, T. (1992). Locally weighted regression in diffuse near-infrared transmittance spectroscopy. *Applied Spectroscopy*, *46*(1), 34-43.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user friendly guide to multivariate calibration and classification*. NIR publications.

Ni, W., Nørgaard, L., & Mørup, M. (2014). Non-linear calibration models for near infrared spectroscopy. *Analytica chimica acta*, *813*, 1-14.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*. Longman scientific and technical.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. *Geoderma*, *195*, 268-279.

Riley, M. R., Rhiel, M., Zhou, X., Arnold, M. A., & Murhammer, D. W. (1997). Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering*, *55*(1), 11-15.

Riley, M. R., Arnold, M. A., & Murhammer, D. W. (1998a). Matrix-enhanced calibration procedure for multivariate calibration models with near-infrared spectra. *Applied spectroscopy*, *52*(10), 1339-1347.

Riley, M. R., Arnold, M. A., Murhammer, D. W., Walls, E. L., & DelaCruz, N. (1998b). Adaptive Calibration Scheme for Quantification of Nutrients and Byproducts in Insect Cell Bioreactors by Near-Infrared Spectroscopy. *Biotechnology progress*, *14*(3), 527-533.

Roychoudhury, P., Harvey, L. M., & McNeil, B. (2006). The potential of mid infrared spectroscopy (MIRS) for real time bioprocess monitoring. *Analytica chimica acta*, *571*(2), 159-166.

Roychoudhury, P., McNeil, B., & Harvey, L. M. (2007). Simultaneous determination of glycerol and clavulanic acid in an antibiotic bioprocess using attenuated total reflectance mid infrared spectroscopy. *Analytica chimica acta*, *585*(2), 246-252.

Sarraguça, M. C., Paulo, A., Alves, M. M., Dias, A. M., Lopes, J. A., & Ferreira, E. C. (2009). Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. *Analytical and bioanalytical chemistry*, *395*(4), 1159-1166.

Scarff, M., Arnold, S. A., Harvey, L. M., & McNeil, B. (2006). Near infrared spectroscopy for bioprocess monitoring and control: current status and future trends. *Critical Reviews in Biotechnology*, *26*(1), 17-39.

Schenk, J., Marison, I. W., & von Stockar, U. (2007). A simple method to monitor and control methanol feeding of Pichia pastoris fermentations using mid-IR spectroscopy. *Journal of biotechnology*, *128*(2), 344-353.

Schenk, J., Balazs, K., Jungo, C., Urfer, J., Wegmann, C., Zocchi, A., Marison, I.W. & von Stockar, U. (2008). Influence of specific growth rate on specific productivity and glycosylation of a recombinant avidin produced by a Pichia pastoris Mut^+ strain. *Biotechnology and bioengineering*, *99*(2), 368-377.

Shaw, A. D., Winson, M. K., Woodward, A. M., McGovern, A. C., Davey, H. M., Kaderbhai, N., ... & Goodacre, R. (1999). Rapid analysis of high-dimensional bioprocesses using multivariate spectroscopies and advanced chemometrics. In *Bioanalysis and Biosensors for Bioprocess Monitoring* (pp. 83-113). Springer Berlin Heidelberg.

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, *5*(4), 223-232.

Sivakesava, S., Irudayaraj, J., & Ali, D. (2001). Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. *Process Biochemistry*, *37*(4), 371-378.

Sirén, N., Weegar, J., Dahlbacka, J., Kalkkinen, N., Fagervik, K., Leisola, M., & Von Weymarn, N. (2006). Production of recombinant HIV-1 Nef (negative factor) protein using Pichia pastoris and a low-temperature fed-batch strategy. *Biotechnology and applied biochemistry*, *44*(3), 151-158.

Soons, Z. I., Streefland, M., van Straten, G., & van Boxtel, A. J. (2008). Assessment of near infrared and "software sensor" for biomass monitoring and control. *Chemometrics and Intelligent Laboratory Systems*, *94*(2), 166-174.

Spanjers, H., Bouvier, J. C., Steenweg, P., Bisschops, I., Van Gils, W., & Versprille, B. (2006). Implementation of in-line infrared monitor in full-scale anaerobic digestion process. *Water science and technology*, *53*(4-5), 55-61.

Steyer, J. P., Bouvier, J. C., Conte, T., Gras, P., Harmand, J., & Delgenes, J. P. (2002). On-line measurements of COD, TOC, VFA, total and partial alkalinity in anaerobic digestion processes using infra-red spectrometry. *Water Science and Technology*, *45*(10), 133-138.

Stuart, B. (2004). *Infrared spectroscopy: fundamentals and applications*. John Wiley & Sons Ltd.

Thomas, E. V., & Haaland, D. M. (1990). Comparison of multivariate calibration methods for quantitative spectral analysis. *Analytical Chemistry*, *62*(10), 1091-1099.

Vaidyanathan, S., McNeil, B., & Macaloney, G. (1999). Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *Analyst*, *124*(2), 157-162.

Vermasvuori, R., Koskinen, J., Salonen, K., Sirén, N., Weegar, J., Dahlbacka, J., ... & von Weymarn, N. (2009). Production of recombinant HIV-1 nef protein using different expression host systems: A techno-economical comparison. *Biotechnology progress*, *25*(1), 95-102.

Wang, Z., Isaksson, T., & Kowalski, B. R. (1994). New approach for distance measurement in locally weighted regression. *Analytical Chemistry*, *66*(2), 249-260.

Ward, A. J., Bruni, E., Lykkegaard, M. K., Feilberg, A., Adamsen, A. P., Jensen, A. P., & Poulsen, A. K. (2011a). Real time monitoring of a biogas digester with gas chromatography, near-infrared spectroscopy, and membrane-inlet mass spectrometry. *Bioresource technology*, *102*(5), 4098-4103.

Ward, A. J., Hobbs, P. J., Holliman, P. J., & Jones, D. L. (2011b). Evaluation of near infrared spectroscopy and software sensor methods for determination of total alkalinity in anaerobic digesters. *Bioresource technology*, *102*(5), 4083-4090.

Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Windig, W., & Koch, R. S. (2006). Chemometrics tutorial for PLS_Toolbox and Solo. *Eigenvector Research Incorporated*, *203.*

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

Yenigün, O., & Demirel, B. (2013). Ammonia inhibition in anaerobic digestion: a review. *Process Biochemistry*, *48*(5), 901-911.

Yeung, K. S., Hoare, M., Thornhill, N. F., Williams, T., & Vaghjiani, J. D. (1999). Near-infrared spectroscopy for bioprocess monitoring and control. *Biotechnology and Bioengineering*, *63*, 684-693

Zhang, M. L., Sheng, G. P., Mu, Y., Li, W. H., Yu, H. Q., Harada, H., & Li, Y. Y. (2009a). Rapid and accurate determination of VFAs and ethanol in the effluent of an anaerobic H 2-producing bioreactor using near-infrared spectroscopy. *Water Research*, *43*(7), 1823-1830.

Zhang, M., Sheng, G., & Yu, H. (2009b). Near-infrared spectroscopy-based quantification of substrate and aqueous products in wastewater anaerobic fermentation processes. *Chinese Science Bulletin*, *54*(11), 1918-1922.

Dissertations published by Process Design and Systems Engineering
ISSN 2489-7272
978-952-12-3682-2
978-952-12-3683-9 (pdf)
Painosalama Oy
Åbo 2018